# Estimation in moderately misspecified models

**Nils Lid Hjort**

*University of Oslo and Norwegian Computing Centre*

-- August 1991 --

ABSTRACT. Suppose data are fitted to some parametric model $f(y,\theta)$, but that the true model happens to be $f(y,\theta,\gamma)$, with an additional $\gamma$ parameter, and where $f(y,\theta,\gamma_0) = f(y,\theta)$. When a parameter $\mu$ is to be estimated, one can use $\mu(\widehat{\theta}_{\rm wide},\widehat{\gamma}_{\rm wide})$ based on likelihood estimation in the wider model, or $\mu(\widehat{\theta}_{\rm narr},\gamma_0)$ based on likelihood estimation of $\theta$ alone in the narrow model. Including $\gamma$ in the model means less bias but larger sampling variability. Two basic questions are addressed in this article. (i) Just how much misspecification can the narrow model tolerate? In the context of a large-sample moderate misspecification framework we find a surprisingly simple, sharp, and general answer, in the form of an explicit criterion for when narrow estimation is more precise than wide estimation. There is effectively a 'tolerance radius' around a given narrow model. The theory is illustrated by computing this tolerance radius in a selection of interesting examples, that also demonstrate the degree of robustness of important standard methods against moderate incorrectness of the model under which they are optimal. (ii) Are there other estimators that work well both under narrow and wide circumstances? We discuss several possibilities and propose some new procedures. All methods are compared in a broad performance study. This comparison can be carried out rather generally and rather simply due to a drastic reduction to a particular standard problem.

KEY WORDS: *Bayes and empirical Bayes, choice of model, compromise estimators, deliberate bias, ignorance is strength, misspecified model, parametric inference, performance study, tolerance radius*

## 1. Introduction and motivating examples

Our theme is moderately misspecified parametric models, and we ask two main questions. The first is: Just how much misspecification can a given parametric model tolerate in a certain direction? More specifically, when is it advantageous to stick to the narrow model, even when it is incorrect? When will 'narrow estimation' be more precise than 'wide estimation'? The second question is broader: Are there estimators that are about as good as the narrow estimator when the narrow model is correct, and about as good as the wide estimator when the narrow model is incorrect? We shall present a generous list of examples to motivate the problems and precise formulations of them.

EXAMPLE A. Suppose data $Y_1,\ldots,Y_n$ come from a life distribution on $[0,\infty)$ and that the median $\mu$ is to be estimated. If the density is the exponential $f(y) = \theta e^{-\theta y}$, then $\mu = \log 2/\theta$, and a natural estimator is $\widehat{\mu}_{\rm narr} = \log 2/\widehat{\theta}_{\rm narr}$, where $\widehat{\theta}_{\rm narr} = 1/\bar{Y}$ is the maximum likelihood (ML) estimator in this narrow model. If it is suspected that the model could deviate from simple exponentiality in direction of the Weibull distribution, with

$$f(y,\theta,\gamma) = \exp\{-(\theta y)^\gamma\}\gamma(\theta y)^{\gamma-1}\theta, \quad y>0, \tag{1.1}$$

then we should conceivably use $\widehat{\mu}_{\rm wide} = (\log 2)^{1/\widehat{\gamma}}/\widehat{\theta}$, using ML estimators $\widehat{\theta}, \widehat{\gamma}$ in the wider Weibull model. But *if* the simple model is right, i.e. $\gamma = 1$, then $\widehat{\mu}_{\rm narr}$ is better, in terms (for example) of mean squared error. By sheer continuity it should be better also for $\gamma$'s close to 1. How much must $\gamma$ differ from 1 in order for $\widehat{\mu}_{\rm wide}$ to become better? And what with similar questions for other typical parametric departures from exponentiality, like the gamma family?

EXAMPLE B. The most popular modelling of data $Y_1, \ldots, Y_n$ is to postulate normality, i.e. assuming $f(y) = \phi((y-\xi)/\sigma)/\sigma$ for suitable parameters $\xi$ and $\sigma$. In many situations the normal density is too light-tailed to constitute a serious description, however. A remedy then is to use

$$f(y, \xi, \sigma, m) = g_m\Big(\frac{y-\xi}{\sigma}\Big)\frac{1}{\sigma},$$

where $g_m(t)$ is the $t$-density with $m$ degrees of freedom. The narrower normal model corresponds to $m = \infty$, and it is naturally felt that for large $m$ the discrepancy between normality and $t$-ness shouldn't matter. Suppose for example that the parameter to be estimated is sd, the standard deviation for $Y_i$'s. For how large $m$ is it the case that the narrow-model estimator $\widehat{\rm sd}_{\rm narr}$, which happens to be the ordinary empirical standard deviation, is better than the more laborious

$$\widehat{\rm sd}_{\rm wide} = \sqrt{\frac{\widehat{m}}{\widehat{m}-2}}\,\widehat{\sigma},$$

computed from ML estimates $\widehat{\xi}$, $\widehat{\sigma}$, $\widehat{m}$ in the three-parameter model? What with other parameters to estimate than the standard deviation?

EXAMPLE C. Consider a regression situation with $n$ pairs $(x_i, Y_i)$. The classical model says $Y_i \sim N\{\alpha + \beta x_i, \sigma^2\}$ for appropriate parameters $\alpha$, $\beta$, $\sigma$, and encourages for example $\widehat{\mu}_{\rm narr} = \widehat{\alpha}_{\rm narr} + \widehat{\beta}_{\rm narr} x$ as the estimator for the median (or mean value) of the distribution of $Y$ for a given $x$ value. Suppose however that the regression curve could be moderately quadratic, say $Y_i \sim N\{\alpha + \beta x_i + \gamma(x_i - \bar{x})^2, \sigma^2\}$ for a moderate $\gamma$. How much must $\gamma$ differ from zero in order for

$$\widehat{\mu}_{\rm wide} = \widehat{\alpha} + \widehat{\beta} x + \widehat{\gamma}(x - \bar{x})^2,$$

with regression parameters now evaluated in the wider model, to perform better? And again the same questions could be asked for other parameters, like comparing $\widehat{x}_{0,\rm narr}$ with $\widehat{x}_{0,\rm wide}$, the narrow-model based and the wide-model based estimators of the point $x_0$ at which the regression curve crosses a certain level.

EXAMPLE D. In some situations a more interesting departure from standard regression lies in variance heterogeneity. This could for example suggest using $Y_i \sim N\{\alpha + \beta x_i, \sigma^2(1 + \gamma x_i)\}$, where $\gamma$ is zero under classical regression. For what range of $\gamma$ values are standard methods, all derived under the $\gamma = 0$ hypothesis, still better than four-parameter-model analysis?

EXAMPLE E. Let us also include another type of model uncertainty, that of misspecification due to using an incorrect transformation. The transformation model invented here

has some of the intentions of the Box–Cox power transformation scheme, but avoids some of its pitfalls. It postulates that

$$h_\lambda\big((Y_i - \alpha - \beta x_i)/\sigma\big) \sim N\{0,1\}, \quad \text{where} \quad h_\lambda(z) = \Phi^{-1}\{\Phi(z)^\lambda\}, \tag{1.2}$$

for appropriate values of $(\alpha, \beta, \sigma, \lambda)$; $\lambda = \lambda_0 = 1$ brings us back to classics. Let us briefly discuss this model and its use before we concentrate on the local misspecification part. It can be written $Y_i = \alpha + \beta x_i + \sigma Z_i$, where $h_\lambda(Z_i)$ follows the standard normal distribution for suitable transformation parameter. Varying $\lambda$ gives a fair range of transformations, and in particular includes the possibility of having skewed error distributions. The four parameters can be estimated from the data. The notation is possibly deceiving, in that it invites one to think in terms of '$\alpha + \beta x_i$ plus noise with level $\sigma$'. This is not quite the case since $Z_i$ has a skewed distribution with mean and median different from zero. It is advisable to reparameterise, after having found a suitable $\lambda$ from data, to the familiar structure + noise form. One possibility is $Y_i = \{\alpha + \sigma e(\lambda)\} + \beta x_i + \sigma v(\lambda) Z_i^0$, in which $e(\lambda)$ and $v(\lambda)$ are mean value and standard deviation of $Z_i$, under the $h_\lambda(Z_i) \sim N\{0,1\}$ model, and where $Z_i^0$ now has mean zero and standard deviation 1. Another possibility is

$$\begin{aligned} Y_i &= \{\alpha + \sigma\Phi^{-1}(0.50^{1/\lambda})\} + \beta x_i + \sigma\{\Phi^{-1}(0.75^{1/\lambda}) - \Phi^{-1}(0.25^{1/\lambda})\} Z_i' \\ &= \alpha' + \beta x_i + \sigma' Z_i', \end{aligned} \tag{1.3}$$

the point being that $Z_i'$ has median zero and interquartile range 1. Our technical point is that (1.2) is a useful generalisation of classical regression to situations with skewed errors, and that parameter estimation is best carried out using ML machinery on (1.2); and our statistical point is that (1.3) better conveys the structure and the noise in the data, and should be used post estimation.

The present concern is how robust standard methods, which presume $\lambda = 1$, are against misspecification of that parameter. Should one use

$$\widehat{\mu}_{\text{wide}}(x) = \widehat{\alpha}_{\text{wide}} + \widehat{\beta}_{\text{wide}} x + \widehat{\sigma}_{\text{wide}} \Phi^{-1}(0.50^{1/\widehat{\lambda}_{\text{wide}}}) \tag{1.4}$$

to estimate the median of $Y$ for given $x$, or will the effortlessly obtainable $\widehat{\mu}_{\text{narr}}(x) = \widehat{\alpha}_{\text{narr}} + \widehat{\beta}_{\text{narr}} x$ suffice?

EXAMPLE F. Next consider logistic regression, in which pairs $(x_i, Y_i)$ are observed of the type $Y_i | x_i \sim \text{Bin}\{1, p(x_i)\}$, with $p(x) = \exp(\alpha + \beta x)/\{1 + \exp(\alpha + \beta x)\}$ being the standard model. Again we can ask whether standard methods based on $(\widehat{\alpha}_{\text{narr}}, \widehat{\beta}_{\text{narr}})$, for example for estimating the true $p(x)$ at a given $x$, or for estimating the cut-off point at which $p(x)$ exceeds $\frac{1}{2}$, become seriously inferior under moderate misspecifications. One natural type of departure is modelled by adding a quadratic term $\gamma(x_i - \bar{x})^2$ to the linear term; another is

$$p(x) = p(x, \alpha, \beta, \eta) = \left\{ \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \right\}^\eta, \tag{1.5}$$

where it is of interest to vary $\eta$ around $\eta_0 = 1$.

EXAMPLE G. Our final example is the two-sample model with variances that may or may not be equal. So $X_1, \ldots, X_m$ are $N\{\xi_1, \sigma_1^2\}$ and $Y_1, \ldots, Y_n$ are $N\{\xi_2, \sigma_2^2\}$, all of them

are independent, and the narrow model specifies that $\sigma_1 = \sigma_2$. Under this assumption it is easy to put up estimators, confidence intervals etc. for parameters related to the difference between the $X$-distribution and the $Y$-distribution, like the Mahalanobis distance $\Delta = |\xi_2 - \xi_1|/\sigma$. More awkward methods are needed when $\sigma_2 \neq \sigma_1$, cf. the Behrens–Fisher problem. The in some sense natural generalisation of the Mahalanobis distance is for example

$$\Delta = (\nu^2 + \omega^2)^{1/2}, \quad \text{where } \nu^2 = (\xi_2 - \xi_1)^2/\sigma^2, \, \omega^2 = 4 \log \frac{\sigma^2}{\sigma_1 \sigma_2}, \, \sigma^2 = (\sigma_1^2 + \sigma_2^2)/2,$$

see Hjort (1986a, Ch. 10). How resistant is the simple $\widehat{\Delta}_{\text{narr}} = |\bar{Y} - \bar{X}|/\widehat{\sigma}_{\text{narr}}$ to differences in $\sigma_1$, $\sigma_2$? When is it necessary to use the much more complicated $\widehat{\Delta}_{\text{wide}}$?

Let us summarise the common characteristics of these situations. There is a narrow and usually simple parametric model which can be fitted to the data, but there is a potential misspecification, which can be ameliorated by its encapsulation in a wider model with one additional parameter. Estimating a parameter assuming correctness of the narrow model involves modelling bias, but doing it in the wider model could mean larger sampling variability. Thus the choice of method becomes a statistical balancing act with perhaps deliberate bias against variance.

The examples above span a reasonable range of heavily used 'narrow' models along with indications of rather typical kinds of deviances from them. Many standard textbook methods for parametric inference are derived under the conditions of such narrow models. Our main result, derived in Section 3, is a surprisingly sharp and general large-sample criterion for how much misspecification a given narrow model can tolerate. This criterion is applied to Examples A–G in Section 7. It is relatively easy to compute, in that it only involves the familiar Fisher information matrix, for the wide model, but evaluated under narrow model conditions. A particularly pleasing facet of our tolerance criterion is that it does not depend upon the particular parameter estimand at all!

In addition to quantifying the degree of robustness of standard methods there are also pragmatic reasons for the present investigation. Statistical analysis will in practice still be carried out using narrow model based methods in the majority of cases, for reasons of ignorance, simplicity, naïvité and boldness; using wide model methods will very often be much more laborious, and only experts will use them anyhow. Thus it is of interest to quantify the consequences of ignorance, and it would be nice to obtain permission to go on doing analysis as if the simple model were true. Such a partial permission is in fact given here. The results of this paper can be interpreted as saying that mild departures from the narrow model do not really matter, and that in fact more ambitious methods could perform worse. In the examples of Section 7 quite explicit limits are given for the degree of misspecification that is tolerable. But this upper limit is in most cases dependent upon parameters of the model, and should be estimated by the conscientious statistician in situations where departures of the type described are suspected. So one should by all means carry out estimation of the additional parameter, even if it turns out that it won't be needed in the final analysis.

Several tangential topics are taken up in Section 4. These include measures of distance from null model to the least tolerable misspecification; comparison with the model selection

criteria of Akaike and Schwarz; simulation based evaluation of our criterion; discussion of the concept of a robust model; dangerous versus noncritical departures from a model; interpretation of confidence intervals under misspecification; and deviances from a model in more than one direction.

There is also room for improvement over the narrow and wide methods. In Section 5 some new estimators are proposed that are designed to work well both under narrow and wide circumstances. A broad comparison of the various compromise estimators is made, in a large-sample framework of moderately misspecified parametric models. A connection to Bayesian robustness is also made. We are able to make a quite general and drastic reduction: The performance of a large class of competing estimators can be studied in a much simpler and very classical context, that of estimating $a$ in a $N\{a,1\}$ situation with one observation! Here the narrow model corresponds to $a = 0$. This provides fresh motivation for studying $a$-estimators that in various ways take into account that values of $a$ in the vicinity of zero are perhaps more likely or perhaps more important. Such a study is reported on in Sections 5 and 6.

The traditional robustness literature is mostly concerned with construction of methods that perform well over a 'nonparametric neighbourhood' around some basic model. The present work is different in that it envisages specific, parametric alternatives to the basic model. There is a literature on parametric robustness, perhaps chiefly concerned with studying behaviour of standard methods and modified standard methods under natural violations of the basic model. Only rarely have comparisons been made between 'narrow' and 'wide' methods, however. Some papers have calculated and commented on the increased estimation noise for a narrow model parameter when passing to a wider model, like comparing the variances of $\widehat{\theta}_{\rm narr}$ and $\widehat{\theta}_{\rm wide}$ in Example A. This is beside the point, partly confusing, and not very interesting, since what matters is studying 'real' parameters which are meaningful functions of the full model, as the median $\mu = \mu(f) = \mu(\theta,\gamma) = (\log 2)^{1/\gamma}/\theta$ in Example A, the standard deviation $\text{sd} = \text{sd}(f) = \text{st}(\xi,\sigma,m) = \{m/(m-2)\}^{1/2}\sigma$ in Example B, etcetera. Bickel (1984) is on the other hand clear about this issue, and is concerned with several problems that resemble those considered here. He does not compare narrow and wide methods, and does not study tolerance distances, but works directly with certain minimax strategies, in a framework of nested linear normal models; see also 5G below. The paper by Berger (1982) on Bayesian robustness also turns out to be related to some of these questions. See Bickel's comments on Berger and 5E, 5F, 5G below.

## 2. Large-sample framework for the problem

We shall start our investigation in the i.i.d. framework. Suppose $Y_1,\ldots,Y_n$ come from some common density $f$, and represent the wide model as $f(y) = f(y,\theta,\gamma)$, where $\gamma = \gamma_0$ corresponds to the narrow model, say $f(y,\theta) = f(y,\theta,\gamma_0)$. We assume that $\theta = (\theta_1,\ldots,\theta_p)'$ lies in some open region in Euclidian $p$-space, that $\gamma$ lies in some interval containing $\gamma_0$, and that the wide model is 'smooth'; for definiteness we postulate that the regularity conditions put forward in Lehmann's (1983) Chapter 6.4 are in force. We are to study behaviour of estimators when $\gamma$ deviates from $\gamma_0$. The parameter to be estimated is some $\mu = \mu(f)$, which we write as $\mu(\theta,\gamma)$ since the wider model is assumed to be an adequate description of reality. We concentrate on ML procedures, and write $\widehat{\theta}_{\rm narr}$ for the

estimator of $\theta$ in the narrow model and $(\widehat{\theta},\widehat{\gamma})$ for the estimators in the wide model. The two major entries in the competition are

$$\widehat{\mu}_{\rm narr} = \mu(\widehat{\theta}_{\rm narr},\gamma_0) \quad \text{and} \quad \widehat{\mu}_{\rm wide} = \mu(\widehat{\theta},\widehat{\gamma}) \tag{2.1}$$

(but see Section 5 for other estimators).

These could be compared in an asymptotic framework in which $Y_i$'s come from some fixed $f(y,\theta_0,\gamma)$, and $\gamma \neq \gamma_0$. In this case $\sqrt{n}(\widehat{\mu}_{\rm wide} - \mu)$ has a limit distribution, which can be derived from the proposition below. The situation is different for the narrow model procedure. Here $\sqrt{n}(\widehat{\mu}_{\rm narr} - \mu)$ can be represented as a sum of two terms. The first is $\sqrt{n}\{\mu(\widehat{\theta}_{\rm narr},\gamma_0) - \mu(\theta_0,\gamma_0)\}$, which has a limit distribution, with generally smaller variability than that of the wide model procedure, and the second is $-\sqrt{n}\{\mu(\theta_0,\gamma) - \mu(\theta_0,\gamma_0)\}$, which tends to plus or minus infinity, reflecting a bias that for very large $n$ will dominate completely. This merely goes to show that with very large sample sizes one is penalised for any bias and one should use the wide model. This result is somewhat irrelevant, however, and suggests that a large-sample framework which uses a local neighbourhood of $\gamma_0$ that shrinks when the sample size grows is much more adequate. Study therefore model $P_n$, the $n$'th model, under which

$$Y_1,\ldots,Y_n \text{ are i.i.d. from } f_n(y) = f(y,\theta_0,\gamma_0+\delta/\sqrt{n}), \tag{2.2}$$

and where $\theta_0$ is fixed but arbitrary. In this framework we need limit distributions for the wide model estimators $(\widehat{\theta},\widehat{\gamma})$ and for the narrow model estimator $\widehat{\theta}_{\rm narr}$.

Consider

$$\begin{pmatrix} U(y) \\ V(y) \end{pmatrix} = \begin{pmatrix} \partial \log f(y,\theta_0,\gamma_0)/\partial\theta \\ \partial \log f(y,\theta_0,\gamma_0)/\partial\gamma \end{pmatrix}, \tag{2.3}$$

the score function for the wide model, but evaluated at the null point $(\theta_0,\gamma_0)$; and the accompanying familiar $(p+1)\times(p+1)$ size information matrix

$$J_{\rm wide} = {\rm VAR}_0 \begin{pmatrix} \partial \log f(Y,\theta_0,\gamma_0)/\partial\theta \\ \partial \log f(Y,\theta_0,\gamma_0)/\partial\gamma \end{pmatrix} = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix}.$$

Note that the $p\times p$ size $J_{11}$ is simply the information matrix of the narrow model, evaluated at $\theta_0$, and that the scalar $J_{22}$ is the variance of $V(Y_i)$, also computed under the narrow model.

PROPOSITION. *Under the sequence of models $P_n$ of (2.2), as $n$ tends to infinity, we have*

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}-\theta_0) \\ \sqrt{n}(\widehat{\gamma}-\gamma_0-\delta/\sqrt{n}) \end{pmatrix} \to_d N_{p+1}\{0, J_{\rm wide}^{-1}\}, \text{ or } \begin{pmatrix} \sqrt{n}(\widehat{\theta}-\theta_0) \\ \sqrt{n}(\widehat{\gamma}-\gamma_0) \end{pmatrix} \to_d N_{p+1}\{\begin{pmatrix} 0 \\ \delta \end{pmatrix}, J_{\rm wide}^{-1}\};$$

$$\sqrt{n}\{\widehat{\theta}_{\rm narr} - (\theta_0 + J_{11}^{-1}J_{12}\delta/\sqrt{n})\} \to_d N_p\{0, J_{11}^{-1}\}, \text{ or } \sqrt{n}(\widehat{\theta}_{\rm narr}-\theta_0) \to_d N_p\{J_{11}^{-1}J_{12}\delta, J_{11}^{-1}\}.$$

PROOF: Consider $\widehat{\theta}_{\rm narr}$ first. The familiar Taylor expansion arguments that lead to the classical $\sqrt{n}(\widehat{\theta}_{\rm narr}-\theta_0)\to_d N\{0,J_{11}^{-1}\}$ under the null model $f(x,\theta_0,\gamma_0)$ can be used in the present $\gamma_0+\delta/\sqrt{n}$ case as well. For

$$\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(Y_i,\widehat{\theta}_{\rm narr},\gamma_0)=\sum_{i=1}^n U(Y_i)+I_n(\widetilde{\theta}_n)(\widehat{\theta}_{\rm narr}-\theta_0)=0,$$

in which $I_n(\theta)=\sum_{i=1}^n \frac{\partial^2}{\partial\theta\partial\theta}\log f(Y_i,\theta,\gamma_0)$ and $\widetilde{\theta}_n$ lies somewhere between $\theta_0$ and $\widehat{\theta}_{\rm narr}$. Under the conditions stated $\widehat{\theta}_{\rm narr}\to_p\theta_0$, under $P_n$, using necessary but moderate variations of the arguments used in Lehmann's (1983) Chapter 6.4 and 6.8, and $-I_n(\theta_0)/n$ as well as $-I_n(\widetilde{\theta}_n)/n$ tend in probability, under $P_n$, to $J_{11}$. All this leads to

$$\sqrt{n}(\widehat{\theta}_{\rm narr}-\theta_0)\doteq_d\{-\tfrac{1}{n}I_n(\theta_0)\}^{-1}\sqrt{n}\bar{U}_n\doteq_d J_{11}^{-1}\sqrt{n}\bar{U}_n, \tag{2.4}$$

where $A_n\doteq_d B_n$ means that $A_n-B_n$ tends to zero in probability, and $\bar{U}_n$ is the average of the $n$ first $U(Y_i)$'s. The triangular version of the Lindeberg theorem shows that $\sqrt{n}\bar{U}_n$ tends in distribution, under $P_n$, to $N_p\{J_{12}\delta,J_{11}\}$. This is because

$$\begin{aligned}E_{P_n}U(Y_i)&=\int f(y,\theta_0,\gamma_0+\delta/\sqrt{n})U(y)\,dx\\&\doteq\int f(y,\theta_0,\gamma_0)\{1+V(y)\delta/\sqrt{n}\}U(y)\,dy=J_{12}\delta/\sqrt{n},\end{aligned}$$

and similarly $U(Y_i)U(Y_i)'$ can be shown to have expected value $J_{11}+O(\delta/\sqrt{n})$, under $P_n$. This proves the 'narrow' part of the proposition.

Similar reasoning takes care of the 'wide' part too. One finds

$$\begin{pmatrix}\sqrt{n}(\widehat{\theta}-\theta_0)\\ \sqrt{n}(\widehat{\gamma}-\gamma_0)\end{pmatrix}\doteq_d J_{\rm wide}^{-1}\begin{pmatrix}\sqrt{n}\bar{U}_n\\ \sqrt{n}\bar{V}_n\end{pmatrix}\to_d J_{\rm wide}^{-1}N_{p+1}\{\begin{pmatrix}J_{12}\delta\\ J_{22}\delta\end{pmatrix},J_{\rm wide}\}, \tag{2.5}$$

which is equivalent to the wide part statement. □

REMARK. Let us for a moment consider more general departures from the $f(y,\theta)$ model. Assume only that data $Y_i$ come from a fixed $f$. Then one can show that $\widehat{\theta}_{\rm narr}$ is consistent for the particular 'least false' or 'most appropriate' parameter value $\theta_{\rm l.f.}=\theta(f)$ that minimises the Kullback–Leibler distance $d[f\colon f(.,\theta)]=\int f(y)\log\{f(y)/f(y,\theta)\}\,dy$. One can also show that $\sqrt{n}(\widehat{\theta}_{\rm narr}-\theta_{\rm l.f.})$ tends in distribution to $N_p\{0,J(f)^{-1}K(f)J(f)^{-1}\}$, in which

$$J(f)=-E_f\frac{\partial^2\log f(Y,\theta_{\rm l.f.})}{\partial\theta\partial\theta},\quad\text{and}\quad K(f)=\mathrm{VAR}_f\frac{\partial\log f(Y,\theta_{\rm l.f.})}{\partial\theta}. \tag{2.6}$$

This is for example made clear in Hjort (1986a, Ch. 3). — Let us apply this to the local misspecification situation, that is, insert $f(y)=f(y,\theta_0,\gamma)$, where $\gamma$ is close to $\gamma_0$. Then, by judicious Taylor expansion arguments, one can show that

$$\theta_{\rm l.f.}=\theta_0+J_{11}^{-1}J_{12}(\gamma-\gamma_0)+O((\gamma-\gamma_0)^2),\quad J(f)^{-1}K(f)J(f)^{-1}=J_{11}^{-1}+O(\gamma-\gamma_0).$$

Using this, for the local $\gamma = \gamma_0 + \delta/\sqrt{n}$, can be used to prove the 'narrow part' of the proposition again. Note that the notion and interpretation of a best fitting parameter changes when the model changes, and that the results about $\theta_{\text{l.f.}}$ quantify this in a precise way.

## 3. Solution

In the large-sample framework of the previous section we are to compare two estimators: The 'safe' $\widehat{\mu}_{\text{wide}} = \mu(\widehat{\theta}, \widehat{\gamma})$ based on ML estimation in the wide model, and the 'risky' $\widehat{\mu}_{\text{narr}} = \mu(\widehat{\theta}_{\text{narr}}, \gamma_0)$. The true parameter is $\mu_{\text{true}} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$ under $P_n$, the $n$'th model.

First consider the safe estimator. By the delta method of linearisation we find

$$\begin{aligned}
&\sqrt{n}\{\mu(\widehat{\theta}, \widehat{\gamma}) - \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})\} \\
&\quad \doteq_d (\tfrac{\partial \mu}{\partial \theta})' \sqrt{n}(\widehat{\theta} - \theta_0) + \{(\tfrac{\partial \mu}{\partial \gamma}) + O(1/\sqrt{n})\}\sqrt{n}(\widehat{\gamma} - (\gamma_0 + \delta/\sqrt{n})) \to_d N\{0, \tau^2\},
\end{aligned}$$

where

$$\tau^2 = \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}' J_{\text{wide}}^{-1} \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}. \tag{3.1}$$

The partial derivatives are computed at the null point $(\theta_0, \gamma_0)$. Similarly, for the risky estimator,

$$\begin{aligned}
&\sqrt{n}\{\mu(\widehat{\theta}_{\text{narr}}, \gamma_0) - \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})\} \\
&\quad = \sqrt{n}\{\mu(\widehat{\theta}_{\text{narr}}, \gamma_0) - \mu(\theta_0, \gamma_0)\} - \sqrt{n}\{\mu(\theta_0, \gamma_0 + \delta/\sqrt{n}) - \mu(\theta_0, \gamma_0)\} \\
&\quad \doteq_d (\tfrac{\partial \mu}{\partial \theta})' \sqrt{n}(\widehat{\theta}_{\text{narr}} - \theta_0) - \sqrt{n}\tfrac{\partial \mu}{\partial \gamma}\delta/\sqrt{n} \to_d N\{b\delta, \tau_0^2\},
\end{aligned}$$

in which

$$b = J_{21} J_{11}^{-1} (\tfrac{\partial \mu}{\partial \theta}) - \tfrac{\partial \mu}{\partial \gamma} \quad \text{and} \quad \tau_0^2 = (\tfrac{\partial \mu}{\partial \theta})' J_{11}^{-1} (\tfrac{\partial \mu}{\partial \theta}). \tag{3.2}$$

By evaluating the mean value of the square of the limit distributions we have that $n$ times the asymptotic mean squared error of $\widehat{\mu}_{\text{wide}}$ becomes $\tau^2$, while the corresponding quantity for $\widehat{\mu}_{\text{narr}}$ becomes $b^2\delta^2 + \tau_0^2$.

We are now in a position to find out when the risky estimator is better than the safe one, simply by algebraically solving the inequality $b^2\delta^2 + \tau_0^2 \leq \tau^2$ w.r.t. $\delta$. Start out writing

$$J_{\text{wide}}^{-1} = \begin{pmatrix} J^{11} & J^{12} \\ J^{21} & J^{22} \end{pmatrix},$$

where a prominent rôle is designated for

$$J^{22} = \kappa^2 = (J_{22} - J_{21} J_{11}^{-1} J_{12})^{-1} \tag{3.3}$$

in what follows, and $J^{12} = -J_{11}^{-1} J_{12} \kappa^2$, $J^{11} = J_{11}^{-1} + J_{11}^{-1} J_{12} J_{21} J_{11}^{-1} \kappa^2$. This leads to the simplification

$$\begin{aligned}
\tau^2 &= (\tfrac{\partial \mu}{\partial \theta})' J_{11}^{-1} (\tfrac{\partial \mu}{\partial \theta}) + (\tfrac{\partial \mu}{\partial \theta})' J_{11}^{-1} J_{12} J_{12}' J_{11}^{-1} (\tfrac{\partial \mu}{\partial \theta}) \kappa^2 - 2(\tfrac{\partial \mu}{\partial \theta})' J_{11}^{-1} J_{12} (\tfrac{\partial \mu}{\partial \gamma}) \kappa^2 + (\tfrac{\partial \mu}{\partial \gamma})^2 \kappa^2 \\
&= \tau_0^2 + b^2 \kappa^2.
\end{aligned}$$

We have reached

RESULT. *(i) The case where $b = 0$ is rather trivial; this typically corresponds to asymptotic independence between $\widehat{\theta}$ and $\widehat{\gamma}$ under the null model, and a parameter $\mu$ functionally independent of $\gamma$. In this case $\widehat{\mu}_{\rm wide}$ and $\widehat{\mu}_{\rm narr}$ are asymptotically equivalent, regardless of $\delta$. (ii) In the more interesting case $b \neq 0$, the narrow model based estimator is better than or as good as the wider model based estimator if and only if $\delta^2 \leq \kappa^2$, or $|\delta| \leq \kappa$, or $|\gamma - \gamma_0| \leq \kappa/\sqrt{n}$.*

*Extension to regression models.* To solve the problems raised in the regression type examples of the introduction we also need the similar result in the more general situation of independent observations with covariates. This can be done in a fairly straightforward fashion. Examples C–F of Sections 1 and 7 lead us naturally to the following general framework. Suppose $(x_i, Y_i)$ are independent pairs, where $Y_i$ has density $f(y_i, \sigma, \beta, \gamma | x_i)$ for given $x_i$-value, carrying some scale parameter $\sigma$ (but not necessarily), a vector $\beta = (\beta_1, \ldots, \beta_p)'$ of ordinary regression parameters, plus some interesting one-dimensional extra parameter $\gamma$ that signals departure from the underlying classical model, which corresponds to some appropriate $\gamma = \gamma_0$.

Under mild regularity conditions the main result above continues to be true for regression models, with $\kappa^2$ defined as in (3.3), but with a somewhat more cumbersome $J_{\rm wide}$ matrix than before. The correct definition is now

$$J_{\rm wide} = \lim_{n\to\infty} J_{n,\rm wide} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} {\rm VAR}_0 \begin{pmatrix} \partial \log f(Y_i, \sigma_0, \beta_0, \gamma_0 | x_i)/\partial\sigma \\ \partial \log f(Y_i, \sigma_0, \beta_0, \gamma_0 | x_i)/\partial\beta \\ \partial \log f(Y_i, \sigma_0, \beta_0, \gamma_0 | x_i)/\partial\gamma \end{pmatrix}, \qquad (3.4)$$

where the variance matrices are computed at the null model, under $(\sigma_0, \beta_0, \gamma_0)$. The necessary regularity conditions can be put up in various forms. These would be Lindebergian to secure normal limits and must in particular imply convergence of $J_{n,\rm wide}$; this usually follows if it is assumed that the collection of $x_i$'s come from some distribution in the design space. In practice one would typically use $J_{n,\rm wide}$ to compute $\kappa^2$. Examples are given in Section 7.

## 4. Discussion

*4A. Simplicity.* It is remarkable that the criterion reached does not depend on the particularities of the specific parameter $\mu(\theta, \gamma)$ at all. Thus, in the situation of Example A in the introduction, calculations in Section 7 show that $|\gamma - 1| \leq 1.245/\sqrt{n}$ guarantees that being simple-minded, assuming exponentiality, works better than being ambitious, using a gamma-family, for *every* smooth parameter $\mu(\theta, \gamma)$.

Our criterion $\delta^2 \leq \kappa^2$ can be evaluated and assessed just from knowledge of $J_{\rm wide}$, the information matrix of the full model, but computed at the narrow model only. This is fortunate, as the general $p+1$ parameter matrix will be very hard to compute in many applications, but will be simpler and manageable at the null model. This is demonstrated in Section 7. Observe that the $|\delta| \leq \kappa$ criterion can be thought of in terms of the limiting variance for $\widehat{\gamma}$, at the null model, since $\sqrt{n}(\widehat{\gamma} - \gamma_0)$ tends to $N\{0, \kappa^2\}$ then.

*4B. How far away is the border line?* We have shown that the simple $\theta$ parameter model can tolerate up to $\gamma_0 + \kappa/\sqrt{n}$ deviation from $\gamma_0$ in the encapsulating $(\theta, \gamma)$ model. How far is the border line $\delta = \kappa$ from the narrow model? One way of answering this is in terms of the probability of actually detecting that the narrow model is wrong. The natural 5% level test for the correctness of the narrow model, against the alternative hypothesis that the additional $\gamma$ parameter must be included, is to reject when $Z_n^2 = n(\widehat{\gamma} - \gamma_0)^2/\widehat{\kappa}^2$ exceeds $1.96^2$, since $Z_n^2$ has a limiting $\chi_1^2$ distribution under the narrow model. Here $\widehat{\kappa}$ is any consistent estimator of $\kappa$, or simply equal to the known value in such cases. The probability that this test detects that $\gamma$ is not equal to $\gamma_0$, when it in fact is equal to $\gamma_0 + \delta/\sqrt{n}$, converges to

$$\text{power}(\delta) = \Pr\{\chi_1^2(\delta^2/\kappa^2) > 1.96^2\}, \tag{4.1}$$

featuring the non-central chi squared with 1 degree of freedom and eccentricity parameter $\delta^2/\kappa^2$. This is a consequence of the proposition proved in Section 2. In particular the approximate power at the border case is equal to 17.0%. We can therefore restate the basic result as follows: Provided the true model deviates so modestly from the narrow model, that the probability of detecting it is 17.0% or less with the natural 5% level test, then the risky estimator is better than the safe estimator. Corresponding other figures for (level, power) are, for illustration, (0.01, 0.057), (0.10, 0.264), (0.20, 0.400), (0.29, 0.500).

*4C. Other distance measures.* Let us present a couple of further measures of the distance from null model to border line misspecification. (i) The Kullback–Leibler distance $d[f(.,\theta_0,\gamma_0): f(.,\theta_0,\gamma_0 + \delta/\sqrt{n})]$ can by clever Taylor expansion arguments be shown to be equal to $\frac{1}{2}\delta^2 J_{22}/n$ plus smaller terms, and in the border case the distance becomes $\kappa^2 J_{22}/2n$. (ii) Next consider the so-called statistical distance or $L_1$-distance between the two neighbouring distributions. It is

$$\int |f(y,\theta_0,\gamma_0 + \delta/\sqrt{n}) - f(y,\theta_0,\gamma_0)|\, dy \doteq \frac{\delta}{\sqrt{n}} \int |V(y)|\, f(y,\theta_0,\gamma_0)\, dy.$$

This distance has a direct probabilistical interpretation. In Example A, for example, the $L_1$-distance from exponentiality to the least tolerable Weibull, becomes about $0.923/\sqrt{n}$. (iii) Finally consider weighted $L_2$-distance $\int (f - f_0)^2/f_0\, dy$. An approximation is seen to be $\delta^2 J_{22}/n$, and the least tolerable distance is $\kappa^2 J_{22}/n$. — Note that these three distance measures are transformation invariant. See also 4F.

*4D. Comparison with Akaike's Information Criterion.* The misspecification problem is related to that of choosing a model. One general method for doing this is to use the information criterion of Akaike. In the present setting one is to compare

$$\text{AIC}_{\text{narr}} = \log L_{\text{max,narr}} - p \quad \text{with} \quad \text{AIC}_{\text{wide}} = \log L_{\text{max,wide}} - (p+1),$$

featuring maximised log likelihoods under respectively the narrow model with $p$ parameters and the wide model with $p+1$ parameters. The method consists of choosing the model with largest observed AIC.

It is instructive to study AIC's behaviour in the framework of this article. Using Taylor expansion, along with techniques and notation as in the proof of the proposition of Section 2, one can show that

$$\mathrm{AIC}_{\mathrm{narr}} \doteq_d \sum_{i=1}^{n} \log f(Y_i, \theta_0, \gamma_0) + n\bar{U}_n' J_{11}^{-1} \bar{U}_n - p,$$

$$\mathrm{AIC}_{\mathrm{wide}} \doteq_d \sum_{i=1}^{n} \log f(Y_i, \theta_0, \gamma_0) + n \begin{pmatrix} \bar{U}_n \\ \bar{V}_n \end{pmatrix}' J_{\mathrm{wide}}^{-1} \begin{pmatrix} \bar{U}_n \\ \bar{V}_n \end{pmatrix} - (p+1).$$

Consequently

$$\begin{aligned} \mathrm{AIC}_{\mathrm{wide}} - \mathrm{AIC}_{\mathrm{narr}} &\doteq_d n \begin{pmatrix} \bar{U}_n \\ \bar{V}_n \end{pmatrix}' J_{\mathrm{wide}}^{-1} \begin{pmatrix} \bar{U}_n \\ \bar{V}_n \end{pmatrix} - n\bar{U}_n' J_{11}^{-1} \bar{U}_n - 1 \\ &= n\big[\bar{U}_n'(J^{11} - J_{11}^{-1})\bar{U}_n + 2\bar{U}_n' J^{12} \bar{V}_n + \bar{V}_n^2 J^{22}\big] - 1 \\ &= n(\bar{V}_n - \bar{U}_n' J_{11}^{-1} J_{12})^2 \kappa^2 - 1 \\ &\to_d N\{-\delta/\kappa^2, 1/\kappa^2\}^2 \kappa^2 - 1 = \chi_1^2(\delta^2/\kappa^2) - 1. \end{aligned} \tag{4.2}$$

The probability that AIC prefers the wide model over the narrow model is therefore approximately $\Pr\{\chi_1^2(\delta^2/\kappa^2) > 1\}$. In particular, if the narrow model is perfect, the probability is 0.317, and in the border-line case suggested by this article, i.e. $\delta = \kappa$, the probability is 0.523. All in all the Akaike method agrees very well with the implicit advice of Section 3.

It is also instructive to see that $\mathrm{AIC}_{\mathrm{wide}} - \mathrm{AIC}_{\mathrm{narr}}$ above is asymptotically equivalent to $Z_n^2 - 1$, where $Z_n = \sqrt{n}(\widehat{\gamma} - \gamma_0)/\widehat{\kappa} \to_d N\{\delta/\kappa, 1\}$. This is the test statistic also discussed in 4B. Akaike's method is also related to the pre-test type strategy discussed in Section 5.

Akaike's criterion has a reputation for overfitting too often, and researchers often use a more stingy criterion due to Schwarz. It penalises with the factor $\frac{1}{2}\log n$ times the number of parameters in the model, i.e. subtracts $(\frac{1}{2}\log n)p$ and $(\frac{1}{2}\log n)(p+1)$ instead. The reasoning above, applied to this alternative criterion, shows that the Schwarz method chooses the narrow model, with probability tending (but slowly) to 1. The alternative model must be at least $\delta(\log n)^{1/2}/\sqrt{n}$ away to interest Schwarz [SIC].

*4E. Evaluation of $\kappa$ through stochastic simulation.* The examples of Section 7 show that it is possible to compute $J_{\mathrm{wide}}$ and $\kappa^2$ explicitly even for somewhat complicated departure models, in effect because the computations only need to be carried out at the null model. In some situations it might be too difficult, however. One way out is then to write down the difficult elements of the $J_{\mathrm{wide}}$ matrix in terms of integrals, involving the null density $f(y, \theta_0, \gamma_0)$ as well as $U(y)$ and $V(y)$, and then carry out numerical integration. This is feasible since only one-dimensional integrals are involved. This method gives a numerical value of $\kappa$ for specified basis point $\theta_0$.

Another way is through stochastic simulation. Several options can be considered. (i) Simulate a large number of $Y_i$'s from the null distribution, and compute score functions $U(Y_i)$ and $V(Y_i)$ along the way (see (2.3)). Then compute empirical covariances and variances to get $J_{\mathrm{wide}}$. (ii) Keep $n$ fixed, simulate $Y_1^*, \ldots, Y_n^*$ from the null density, at some desired $\theta_0$, and compute the estimates $\widehat{\theta}^*$ and $\widehat{\gamma}^*$ based on this pseudo-sample. Do

this a large number of times, and the empirical covariance matrix for $(\widehat{\theta}^*, \widehat{\gamma}^*)$ is $J_{\text{wide}}^{-1}/n$. (iii) Or drop $\widehat{\theta}^*$ and just evalute the empirical standard deviation of $\sqrt{n}(\widehat{\gamma}^* - \gamma_0)$, which is $\kappa$. This is a feasible approach in complex regression models, or in parametric and semiparametric survival data models with censoring, where analytical expressions for $\kappa^2$ cannot be found.

*4F. Good models and dangerous departures.* Which departures from a given narrow model are dangerous, and which are insignificant? And what qualities should a 'good and robust' model have?

We have demonstrated that the narrow model can tolerate $\delta = \sqrt{n}(\gamma - \gamma_0)$ up to the limit $\kappa$ in absolute value. The numerical value of $\kappa$ depends on the scale used, however. The appropriate scale invariant tolerance measure is $d = \kappa^2 J_{22} = J^{22} J_{22}$, as is also suggested by the distances considered in 4C. Two numbers of this kind can be directly compared for two specifically envisaged model departures. A model departure with large $d$ is less dangerous than one with small $d$.

A model deviance can be studied in terms of $V(y) = \partial \log f(y, \theta_0, \gamma_0)/\partial\gamma$, see (2.3). How well is $V(y)$ explained by the existing model, represented by $U(y)$? A natural measure is the so-called maximal correlation, $\rho^2\{U, V\}$, the maximal value of $\text{corr}\{a_1 U_1 + \cdots + a_p U_p, V\}^2$ as $a = (a_1, \ldots, a_p)'$ varies. It is well known and just a piece of linear algebra to prove that $a_0 = J_{11}^{-1} J_{12}$ maximises, with resulting

$$\rho^2\{U, V\} = J_{12}' J_{11}^{-1} J_{12} = 1 - 1/(\kappa^2 J_{22}) = 1 - 1/d. \tag{4.3}$$

This invites a geometrical interpretation for the tolerance limit $d$. The smallest possible value for $d$ is 1, which happens when the model departure is 'completely new' and orthogonal to the existing model, with $J_{12} = 0$. Only a mild departure in this direction can be tolerated. So a dangerous departure is one that can be realistically suspected, in the first place, and which has a small $d$, or a small correlation. A non-critical departure is one that has a large tolerance $d$, or a large correlation, or one that perhaps is unrealistic a priori. — A good and robust model, therefore, is one where realistically suspected deviances have large tolerances $d$. See the examples of Section 7.

*4G. Can we de-bias?* We have demonstrated that narrow estimation, which means introducing a deliberate bias to reduce variability, leads to better estimator precision in a certain radius around the narrow model. The precise quantitative result is that $\sqrt{n}(\widehat{\mu}_{\text{narr}} - \mu_{\text{true}})$ tends to $N\{b\delta, \tau_0^2\}$, see Section 3. Can we remove the bias and do even better?

About the best we can do in this direction is to use $\widehat{\mu}_{\text{db}} = \widehat{\mu}_{\text{narr}} - b(\widehat{\gamma} - \gamma_0)$. Analysis reveals, working from the basis result (5.2) of the next section, that $\sqrt{n}(\widehat{\mu}_{\text{db}} - \mu_{\text{true}})$ tends to $N\{0, \tau_0^2 + b^2\kappa^2\}$. So the bias can be removed, but the price one pays amounts exactly to what was won by deliberate biasing in the first place, and the de-biased estimator is equivalent to $\widehat{\mu}_{\text{wide}}$. The reason for the extra variability is that no consistent estimator exists for $\delta$.

*4H. Dwindling confidence.* We have established that $\widehat{\mu}_{\text{narr}}$ has higher precision than $\widehat{\mu}_{\text{wide}}$ for moderate misspecifications of the narrow model. But what with further inference?

Consider confidence intervals. The usual approximate 90% interval for $\mu$ based on narrow model assumptions is $\text{CI}_{\text{narr}} = \widehat{\mu}_{\text{narr}} \pm 1.645\,\widehat{\tau}_0/\sqrt{n}$, where $\widehat{\tau}_0$ is consistent for $\tau_0$

of (3.2). But in the present local misspecification framework $\sqrt{n}(\widehat{\mu}_{\rm narr} - \mu_{\rm true})$ tends to $N\{b\delta, \tau_0^2\}$, and the bias destroys the 90% property. The probability that $\mu_{\rm true}$ is covered by $\mathrm{CI}_{\rm narr}$ converges to $\Pr[-1.645 \leq N\{b\delta/\tau_0, 1\} \leq 1.645]$. This is always *strictly less than 90%*, unless the narrow model is exactly true or $b$ of (3.2) is zero. Yes, I am shocked. The difference is not necessarily dramatic, in that the coverage probability is above 85% when $|b\delta|/\tau_0$ is smaller than 0.54 and above 80% when the ratio is smaller than 0.77. What is important is that the narrow model based interval always underestimates the confidence, under any model departure from any given parametric model, and that we have an illuminating explicit formula for the true (asymptotic) coverage probability.

It is not possible to remove the bias and still get a shorter honest 90% interval than $\mathrm{CI}_{\rm wide} = \widehat{\mu}_{\rm wide} \pm 1.645\,\widehat{\tau}_{\rm wide}/\sqrt{n}$. This follows from analysis similar to that in 4G. Thus, in a way, within the chosen large-sample framework, and provided we insist on guaranteed levels, we cannot carry out confidence and testing analysis better than with wide model methods, despite the fact that narrow estimators often have better precision than wide ones. A practical proposal is to use $\widehat{\mu}_{\rm narr}$ when theory and analysis suggest that it is more precise than $\widehat{\mu}_{\rm wide}$, but to supplement it with a confidence interval obtained through nonparametric or wide-model-parametric bootstrapping. The point is to obtain an honest 90% interval, for example, built around $\widehat{\mu}_{\rm narr}$.

Let us finally point out that narrow based intervals in some natural ways perform better than wide model ones, under mild misspecifications, since they are, indeed, narrower. Assume the loss incurred by using CI to cover $\mu$ is of the form

$$L\{(\theta,\gamma), \mathrm{CI}\} = I\{\mu(\theta,\gamma) \notin \mathrm{CI}\} + \sqrt{n}w\,\mathrm{length(CI)},$$

where $w$ is an appropriately chosen weight factor. The idea is to combine the two desiderata of confidence intervals into one measure; they should miss with low probability and have short length. The asymptotic risk functions for $\mathrm{CI}_{\rm narr} = \widehat{\mu}_{\rm narr} \pm z_0\widehat{\tau}_0/\sqrt{n}$ and $\mathrm{CI}_{\rm wide} = \widehat{\mu}_{\rm wide} \pm z_1\widehat{\tau}/\sqrt{n}$, under model $P_n$, become

$$\begin{aligned} \mathrm{risk}_{\rm narr} &= \Pr\big[|N\{b\delta/\tau_0, 1\}| \geq z_0\big] + 2wz_0\tau_0, \\ \mathrm{risk}_{\rm wide} &= \Pr\big[|N\{0,1\}| \geq z_1\big] + 2wz_1(\tau_0^2 + b^2\kappa^2)^{1/2}. \end{aligned} \tag{4.4}$$

Again the best narrow method will be better than the best wide method, for moderate deviances $\delta$ from zero.

*4I. Deviances in several directions.* It could be worthwile to study two types of model departure simultaneously, like both quadraticity and variance heterogeneity in regression. Generalising our results to such a situation is not difficult. Suppose $f(y, \theta_0, \gamma_0 + \delta/\sqrt{n})$ is the true model, where $\gamma = (\gamma_1, \gamma_2)'$ and $\delta = (\delta_1, \delta_2)'$ are two-dimensional. The natural criterion for when each narrow estimator is asymptotically more precise than its wide contender becomes $\delta\delta' \leq J^{22}$, where $J^{22} = (J_{22} - J_{12}'J_{11}^{-1}J_{12})^{-1}$ is $2 \times 2$. This describes an ellipse around the null model. The border line, the crossing of which means coming into wide supremacy territory, is about $\mathrm{Tr}(J_{22}J^{22})/2n$ away, as measured by Kullback–Leibler. The power of the 5% level $Z_n^2 = n(\widehat{\gamma} - \gamma_0)'(\widehat{J}^{22})^{-1}(\widehat{\gamma} - \gamma_0)$ test is 33.4% at the border.

Most of the general results of Section 5 can similarly be extended to the situation with more than one type of deviance present. See remark (v) of 5H.

## 5. Classes of compromise estimators

We have so far concentrated on $\widehat{\mu}_{\rm narr}$ and $\widehat{\mu}_{\rm wide}$ to estimate $\mu = \mu_{\rm true} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$. These rather cyclopic estimators can however be combined to form dimeric ones that perhaps work well both under the null model and the local alternative. This section considers and develops various more complex estimators with this aim. Some key words indicating the different types that will be discussed are pre-test or if-else estimators, mixture or weighted estimators, Bayes and empirical Bayes estimators, minimax estimators, the Bayesian epsilon estimator, and limited translation estimators.

Comparing all of these approaches may appear to be a formidable task, since the problem conceivably depends upon the particularities of the narrow model, the type and degree of deviance from it, and the specific parameter estimand under study. The comparison problem can however be drastically reduced, as we show in 5D below. Each of a large class of estimators for $\mu_{\rm true}$ has a cousin which estimates $a$ in a $N\{a,1\}$ situation with one observation under squared error loss! The underlying one-one correpondence makes it possible to study the performance of general estimation approaches rather simply and rather generally, and this is indeed done in Section 6.

*5A. If-else of pre-test estimators.* 'The responsibility of tolerance lies with those who have the wider vision.' A procedure that is sometimes advocated in model choice problems and which perhaps is consistent with George Eliot's view is as follows, in the present context: Test the hypothesis $\gamma = \gamma_0$ against the alternative $\gamma \neq \gamma_0$, say at the 10% level; if accepted, then use $\widehat{\mu}_{\rm narr}$, if rejected, then use $\widehat{\mu}_{\rm wide}$. Choosing the $Z_n^2 = n(\widehat{\gamma} - \gamma_0)^2/\widehat{\kappa}^2$ test also discussed in 4B, this suggestion amounts to

$$\widehat{\mu}_{\rm pre} = \widehat{\mu}_{\rm narr} I\{Z_n^2 \leq 1.645^2\} + \widehat{\mu}_{\rm wide} I\{Z_n^2 > 1.645^2\}, \quad 1.645^2 = \text{upper 10\% point of } \chi_1^2. \tag{5.1}$$

But this method sticks too rigidly to the narrow model. The theory of Section 3 suggests that one should rather use the much smaller value 1 as cut-off point, since $|\delta| \leq \kappa$ corresponds to $n(\gamma - \gamma_0)^2/\kappa^2 \leq 1$, and $Z_n^2$ estimates this ratio. Using 1 as cut-off is asymptotically the same as the Akaike model choice strategy, see 4D, and corresponds to a much more relaxed significance level, indeed to 31.7%, which in a way becomes the optimally chosen significance level. Observe that $\sqrt{n}(\widehat{\mu}_{\rm pre} - \mu_{\rm true})$ tends to a mixture of two normals, as further commented upon below.

*5B. Mixture estimators.* Another natural idea is $\widehat{\mu}_{\rm lin} = (1-c)\widehat{\mu}_{\rm narr} + c\widehat{\mu}_{\rm wide}$. To find the approximate distribution of this estimator it is necessary to go somewhat beyond the basic proposition of Section 2, in that the simultaneous limit distribution of the narrow and the wide estimators is needed. This can be found by studying the proof of the proposition, however. Utilising (2.4) and (2.5) it follows via some analysis that

$$\begin{aligned}
\sqrt{n}(\widehat{\mu}_{\rm narr} - \mu_{\rm true}) &\to_d b\delta + (\tfrac{\partial \mu}{\partial \theta})' J_{11}^{-1} M, \\
\sqrt{n}(\widehat{\mu}_{\rm wide} - \mu_{\rm true}) &\to_d \begin{pmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{pmatrix}' J_{\rm wide}^{-1} \begin{pmatrix} M \\ N \end{pmatrix}, \\
Z_n = \sqrt{n}(\widehat{\gamma} - \gamma_0)/\widehat{\kappa} &\to_d Z = (\delta + J^{21} M + J^{22} N)/\kappa,
\end{aligned} \tag{5.2}$$

in which $(M,N) \sim N_{p+1}\{0, J_{\rm wide}\}$. The convergence is simultaneous, and takes place under the $P_n$ sequence of models (2.2). Note that $Z \sim N\{\delta/\kappa, 1\}$.

Now the limit distribution of $\widehat{\mu}_{\rm lin}$ can be obtained. The result is

$$\sqrt{n}(\widehat{\mu}_{\rm lin} - \mu_{\rm true}) \to_d (1-c)b\delta + (1-c)(\tfrac{\partial\mu}{\partial\theta})' J_{11}^{-1} M + c \begin{pmatrix} \frac{\partial\mu}{\partial\theta} \\ \frac{\partial\mu}{\partial\gamma} \end{pmatrix}' J_{\rm wide}^{-1} \begin{pmatrix} M \\ N \end{pmatrix}.$$

This is a normal distribution with mean value $(1-c)b\delta$, and its variance can be shown to be $\tau_0^2 + c^2b^2\kappa^2$, in the notation of Section 3. The ideal value of $c$ that minimises the asymptotic mean squared error for $\widehat{\mu}_{\rm lin}$ is $c_0 = \delta^2/(\kappa^2+\delta^2) = a^2/(1+a^2)$, featuring the key quantity $a = \delta/\kappa$. The accompanying minimum value is equal to $b^2\kappa^2a^2/(1+a^2) + \tau_0^2$. Note that this is always better than both the $b^2\kappa^2 + \tau_0^2$ achieved by $\widehat{\mu}_{\rm wide}$ and the $b^2\delta^2 + \tau_0^2$ achieved by $\widehat{\mu}_{\rm narr}$.

The problem is of course that $c_0$ is unknown since $\delta$ is. Using the empirical counterpart of $\delta = \sqrt{n}(\gamma - \gamma_0)$ invites $Z_n = \sqrt{n}(\widehat{\gamma} - \gamma_0)/\widehat{\kappa}$ to be inserted for $\delta/\kappa$, i.e. $Z_n^2$ estimates $a^2$, and one could try out the diophthalm

$$\widehat{\mu}_{\rm eb} = \frac{1}{1+Z_n^2}\widehat{\mu}_{\rm narr} + \frac{Z_n^2}{1+Z_n^2}\widehat{\mu}_{\rm wide}. \tag{5.3}$$

Note the Steinean overtones. The empirical Bayes connection that gives its subscript is noted in 5F below.

*5C. Compromise estimators.* Let us generalise. We shall be content to study estimators in the fairly large class of *compromise estimators*, which are bilingual and want the best from two worlds, and which we describe as follows. Its prime members are of the type

$$\mu^* = \{1 - c(Z_n)\}\widehat{\mu}_{\rm narr} + c(Z_n)\widehat{\mu}_{\rm wide}, \tag{5.4}$$

where $c(z)$ is almost everywhere continuous. Note that the previously considered estimators are of this form. The additional members that are included are those that can be closely approximated by (5.4) type ones by linearisation. More specifically, the limit distribution result (5.5) below is required to hold. It suffices for $\mu^*$ to be of the form $m(\widehat{\mu}_{\rm narr}, \widehat{\mu}_{\rm wide}, Z_n)$ for some smooth function $m(\mu_1,\mu_2,z)$ with the property that $m(\mu,\mu,z) \equiv \mu$. An example is the harmonic variety $\exp\big[\{1-h(Z_n)\}\log\widehat{\mu}_{\rm narr} + h(Z_n)\log\widehat{\mu}_{\rm wide}\big]$ (which can be used in cases where $\mu$ is positive).

*5D. Comparison of estimators: a drastic reduction.* We wish to study the performance of all these estimators, and to compare pairs of them, w.r.t. the limiting mean squared error criterion.

THEOREM. *The compromise estimator (5.4) has limit distribution, under $P_n$ of (2.2), given by*

$$\sqrt{n}(\mu^* - \mu_{\rm true}) \to_d \Lambda = \{1-c(Z)\}\{b\delta + (\tfrac{\partial\mu}{\partial\theta})' J_{11}^{-1} M\} + c(Z) \begin{pmatrix} \frac{\partial\mu}{\partial\theta} \\ \frac{\partial\mu}{\partial\gamma} \end{pmatrix}' J_{\rm wide}^{-1} \begin{pmatrix} M \\ N \end{pmatrix}. \tag{5.5}$$

*The mean squared error of the limit distribution can be written as*

$$E\Lambda^2 = b^2\kappa^2\, E\big\{\delta/\kappa - c(Z)Z\big\}^2 + \tau_0^2 = b^2\kappa^2 R(\delta/\kappa) + \tau_0^2, \tag{5.6}$$

*in which*

$$R(a) = E\big\{c(Z)Z - a\big\}^2 \text{ and } Z \sim N\{a,1\}. \tag{5.7}$$

PROOF: (5.5) follows from (5.2) and the continuous mapping theorem of weak convergence. To characterise this limit variable $\Lambda$, study its distribution conditional on $Z = z$. Ordinary techniques from multivariate analysis, working from (5.2), lead to

$$\begin{pmatrix} M \\ N \end{pmatrix} | \{Z = z\} \sim N_{p+1}\{\begin{pmatrix} 0 \\ (\kappa z - \delta)/\kappa^2 \end{pmatrix}, \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} - 1/\kappa^2 \end{pmatrix}\}.$$

Several algebraic and multivariate details later one arrives at

$$\Lambda | \{Z = z\} \sim N\{b\delta - c(z)b\kappa z, \tau_0^2\}, \quad \text{where } Z \sim N\{\delta/\kappa, 1\}.$$

Expression (5.6) for the limiting mean squared error can now be worked out, studying first the $z$-conditional and then the unconditional mean value of $\Lambda^2$. □

This result contains those associated with (3.1) and (3.2) as well as (5.1) and the case of fixed $c$ studied above. Observe that the unconditional distribution of $\Lambda$ is non-normal unless $c(z)$ is constant in $z$. Note also that the unfamiliar type of limit distribution is not a peculiarity of the chosen local neighbourhood asymptotics, since $\Lambda$ is typically non-normal even in the null model case.

A particular consequence of the theorem is that *it suffices to compare different versions of the function* $R(a)$, as a function of $a = \delta/\kappa$, since $b\kappa$ and $\tau_0$ remain unchanged for different estimators. [We disregard the rather simple cases in which $b = 0$, see 'case (i)' of Section 3's Result, under which all compromise estimators have $N\{0, \tau_0^2\}$ as limit distribution.] This constitutes an impressive reduction of the original comparison problem. Note that $R(a)$ is simply the risk function for the estimator $c(Z)Z$ for $a$ in the one-observation $Z \sim N\{a, 1\}$ problem under squared error loss. There is a simple one-to-one correspondence from general compromise estimators to estimators $\widehat{a}(Z)$ of $a$ based on $Z$, via

$$\widehat{a}(z) = c(z)z, \quad c(z) = \widehat{a}(z)/z. \tag{5.8}$$

We stress the generality: A comparison between the four natural estimators $\widehat{\mu}_{\text{narr}}$, $\widehat{\mu}_{\text{wide}}$, $\widehat{\mu}_{\text{pre}}$ of (5.1), and $\widehat{\mu}_{\text{eb}}$ of (5.3), for example, can be carried out entirely in the realm of the classical $Z \sim N\{a, 1\}$ situation, by simply drawing the four $R(a)$ curves. See Section 6 for examples. And the conclusions from this comparison remain correct and relevant in *every* 'moderate misspecification' problem, cf. the wide span of problems that Examples A–G represent. Finally one is allowed to go the other way: *Your favourite estimator for* $a$ in the $N\{a, 1\}$ problem (where $a$ may be rumoured to be in the vicinity of zero) can be transported to a useful estimator for any given estimand in any given moderate misspecification situation.

In most cases it holds that $c(z) = c(-z)$, implying $R(a) = R(-a)$, making it necessary to study only non-negative $a$'s. The parameter $a$ measures the degree of misspecification from the narrow model. The important range is perhaps $[-4, 4]$, where $a = 0$ means correctness of the narrow model, $a = \pm 1$ are the turning points after which the wide estimator becomes better than the narrow one, and values beyond $\pm 3$ could be thought of as clearly detectable departures from the narrow model, cf. power considerations (4.1), (4.2). [The 5% level test has power .851 and .979 at $a = 3$ and $a = 4$, whereas the 10% level test has .912 and .991 at the same points.]

These remarks also illustrate the importance of thinking about prior information related to the parameter $a$, for example its possible range. In Example B, studied in Sections 1 and 7 and in Hjort (1991a), $a$ must be non-negative a priori, and in other cases it could be natural to restrict attention to the $[-4,4]$ range, say, or to postulate a prior density for $a$. Such a prior could reflect serious prior beliefs, in the Bayesian fashion, or be used as a mathematical device to reach an estimator with minimum possible averaged mean squared error. Objectivists fretting at such ideas should note that the two classical solutions here, $\widehat{\mu}_{\rm narr}$ and $\widehat{\mu}_{\rm wide}$, correspond to full faith in the priors $I_0$ and 1, respectively, where $I_0$ is the degenerate distribution at zero and 1 is the flat non-informative prior for $a$. This is made clear in the course of the two following subsections, where the correspondence between moderate misspecification problems and the $N\{a,1\}$ situation is explained also for Bayesian matters.

*5E. Prior and posterior distributions for a.* One is used to seeing that 'the prior is washed out by the data'. Assume for example that a prior density $p_0(\theta,\gamma)$ is placed on $(\theta,\gamma)$, with resulting Bayes estimators $(\widehat{\theta}_B,\widehat{\gamma}_B)$, expected values in the posterior density $p_0(\theta,\gamma|\text{data})$. Then these are typically asymptotically equivalent to the ML estimators, in the precise sense that $\sqrt{n}(\widehat{\theta}-\widehat{\theta}_B)\to_p 0$ and $\sqrt{n}(\widehat{\gamma}-\widehat{\gamma}_B)\to_p 0$, in the frequentist framework $P_n$. This is a fairly standard result under null model conditions, and the more delicate case of $\delta\neq 0$ can be treated using methods in Hjort (1986b).

This result uses a fixed prior for $(\theta,\gamma)$, and is somewhat irrelevant in the present context of moderate misspecification. It appears more natural to operate with a fixed prior for $(\theta,\delta)=(\theta,\sqrt{n}(\gamma-\gamma_0))$, or, equivalently, a fixed prior $p(\theta,a)$ for $(\theta,a)=(\theta,\sqrt{n}(\gamma-\gamma_0)/\kappa)$. We think of the prior distribution for $a$ as reflecting prior beliefs about the suitability of the narrow $f(y,\theta,\gamma_0)$ model, cf. the discussion above.

In this situation the prior information regarding $\theta$ will still be overwhelmed by the data, but not the part related to $a$. Information about $a$ lies in $Z_n=\sqrt{n}(\widehat{\gamma}-\gamma_0)/\widehat{\kappa}$, which is not consistent, but has a limiting variable $Z\sim N\{a,1\}$. Intuitively, therefore, the posterior density $p(a|\text{data})$ should for large $n$ simply be close to $p(a|z)$ in the situation where $Z$ is $N\{a,1\}$ and $a$ has prior proportional to $p(\theta_0,a)$. To prove it, let us study $p(a|Y_1,\ldots,Y_n)$ when $n$ grows, under the $P_n$ model, where $f(y)=f(y,\theta_0,\gamma_0+\delta_0/\sqrt{n})$ for some fixed values of $\theta_0$, $\delta_0$. Let $L_n(\theta,\gamma)=\prod_{i=1}^n f(Y_i,\theta,\gamma)$ be the $n$'th likelihood. By judicious second order Taylor expansion analysis it can be established that

$$H_n(s,t)=\frac{L_n(\widehat{\theta}+s/\sqrt{n},\widehat{\gamma}+t/\sqrt{n})}{L_n(\widehat{\theta},\widehat{\gamma})}\to_d H(s,t)=\exp\Big\{-\frac{1}{2}\begin{pmatrix}s\\t\end{pmatrix}'J_{\rm wide}\begin{pmatrix}s\\t\end{pmatrix}\Big\}$$

under $P_n$ of (2.2). The convergence takes place in each Скороход space $D[-A,A]^{p+1}$. Let now $g(\theta,a)$ be any bounded function. Then one may deduce

$$\begin{aligned}E\{g(\theta,a)|\text{data}\}&=\frac{\int\int g(\theta,\sqrt{n}(\gamma-\gamma_0)/\kappa)L_n(\theta,\gamma)p(\theta,\sqrt{n}(\gamma-\gamma_0)/\kappa)\sqrt{n}/\kappa\,d\theta\,d\gamma}{\int\int L_n(\theta,\gamma)p(\theta,\sqrt{n}(\gamma-\gamma_0)/\kappa)\sqrt{n}/\kappa\,d\theta\,d\gamma}\\&=\frac{\int\int g(\widehat{\theta}+s/\sqrt{n},Z_n'+t/\kappa)H_n(s,t)p(\widehat{\theta}+s/\sqrt{n},Z_n'+t/\kappa)\,ds\,dt}{\int\int H_n(s,t)p(\widehat{\theta}+s/\sqrt{n},Z_n'+t/\kappa)\,ds\,dt}\end{aligned}$$

$$\to_d \frac{\int\int g(\theta_0, Z+t/\kappa)H(s,t)p(\theta_0, Z+t/\kappa)\,ds\,dt}{\int\int H(s,t)p(\theta_0, Z+t/\kappa)\,ds\,dt}$$
$$= \frac{\int g(\theta_0, Z+t/\kappa)\exp(-\frac{1}{2}t^2/\kappa^2)\pi(Z+t/\kappa)\,dt}{\int \exp(-\frac{1}{2}t^2/\kappa^2)\pi(Z+t/\kappa)\,dt}$$
$$= \frac{\int g(\theta_0, a)\exp\{-\frac{1}{2}(Z-a)^2\}\pi(a)\,da}{\int \exp\{-\frac{1}{2}(Z-a)^2\}\pi(a)\,da},$$

in which $\pi(a) = \text{const.}\, p(\theta_0, \delta)$ is the prior for $a$ given the information $\theta = \theta_0$, and $Z_n' = (\widehat{\kappa}/\kappa)Z_n$ was used for notational simplicity. The necessary mathematical details have to do with (i) securing convergence inside $[-A, A]^{p+1}$, utilising the proposition of Section 2, along with (5.2); (ii) cleverly carrying out a certain inner $p$-dimensional normal integration; and (iii) bounding integrands outside $[-A, A]^{p+1}$ for large $A$. The arguments that are needed resemble those explained in Hjort (1986b) (to reach a different conclusion, in a different problem), and are left out here.

By considering $g = g(a)$ above it is clear that

$$\pi_n(a|\text{data}) \to_d \pi(a|Z) = \frac{\phi(Z-a)\pi(a)}{\int \phi(Z-a)\pi(a)\,da}, \tag{5.9}$$

under $P_n$, where $Z \sim N\{a, 1\}$ is as in (5.2). This is what was predicted above. If $a$ has some prior distribution $d\pi(a)$ that perhaps does not have a density, then the arguments can be repeated to give $d\pi_n(a|\text{data}) \to_d \text{const.}\, \phi(Z-a)\,d\pi(a)$ instead.

*5F. Bayes and empirical Bayes estimators.* We should distinguish between kosher Bayes and approximate Bayes estimators. A prior density $p(\theta, a)$ for $(\theta, a)$ leads to the exact Bayes solution $\widehat{a}_n = E\{a|Y_1, \ldots, Y_n\}$. This is usually a very complicated expression, and in view of (5.9) it is tempting to work directly in the limit distribution and use $\widehat{a}(Z_n)$ instead, where

$$\widehat{a}(z) = E\{a|Z=z\} = \frac{\int a\phi(z-a)\pi(a)\,da}{\int \phi(z-a)\pi(a)\,da} = z + \frac{\partial}{\partial z}\log\int \phi(z-a)\pi(a)\,da. \tag{5.10}$$

But the arguments of 5E can be used to reach

$$\widehat{a}_n = Z_n' + \frac{\partial}{\partial z}\log\int \phi(Z_n'-a)\pi(a)\,da + O_p(1/\sqrt{n}),$$

where again $Z_n' = (\widehat{\kappa}/\kappa)Z_n$. This proves $\widehat{a}_n - \widehat{a}(Z_n) \to_p 0$, under $P_n$, allowing us to use $\widehat{a}(Z_n)$ instead when we devise and study Bayes solutions in our large-sample framework. In particular we do not have to bother with the part of the prior information that has to do with $\theta$.

Some specific Bayesian and empirical Bayesian constructions follow.

(i) Suppose $a$ is $N\{0, \sigma^2\}$ (where the size of the spread parameter $\sigma$ matters more than the normality). Then $Z \sim N\{0, \sigma^2+1\}$, and $\widehat{a}(z) = \{\sigma^2/(\sigma^2+1)\}\,z$, with Bayes risk

$\sigma^2/(\sigma^2+1)$. If $Ea^2=\sigma^2$ is unknown, a simple guess is $Z_n^2$, since $Z_n$ estimates $a$. This brings forward the empirical Bayes estimate $\widehat{a}_{\rm eb}(Z_n)=\{Z_n^2/(Z_n^2+1)\}\,Z_n$ for $a$. But this corresponds to $\widehat{\mu}_{\rm eb}$ of (5.3), explaining its empirical Bayes interpretation.

One may also consider other estimators for $q=\sigma^2/(\sigma^2+1)$ here. Each such $\widehat{q}=\widehat{q}(Z)$ leads to an $a^*=\widehat{a}(Z,\widehat{q})$, and in its turn to a new estimator $\mu^*$ for $\mu_{\rm true}$ via (5.4) and (5.8). The fact that $EZ^2=\sigma^2+1$ suggests $\widetilde{q}=(Z^2-1)_+/Z^2$, which in fact is the ML solution, or similar versions. Another proposal is to put a vague hyper prior on $\sigma$, or directly on the ratio $q$. The Bayes solution becomes $E\{a|Z=z\}=E_zE_z\{a|\sigma\}=E_z(qz)=\widehat{q}(z)z$, in which $\widehat{q}(z)=\int_0^1 qp(q|z)\,dq$ is the posterior density of $q$ for given $Z=z$. The usual choice for a non-informative prior for a scale parameter like $\sigma$ is to have $\log\sigma$ uniform. This leads to $p(q)=\text{const.}\{q(1-q)\}^{-1}$ on $[\varepsilon,1-\varepsilon]$, say, for $q$, with a corresponding explicit $\widehat{q}(z)$. In fact it turns out that

$$\widehat{q}(z)=\frac{\int_\varepsilon^1(1-q)^{-1/2}\exp\{-\frac{1}{2}(1-q)z^2\}\,dq}{\int_\varepsilon^1 q^{-1}(1-q)^{-1/2}\exp\{-\frac{1}{2}(1-q)z^2\}\,dq} \tag{5.11}$$

is substantially better than $\widetilde{q}=(z^2-l)_+/(z^2+1-l)$, where $0\le l\le 1$, for a wide right interval $(q_0,1)$ of $q$ values. Heroic numerical integrations have demonstrated this, via computions and comparisons of $E_q|q^*-q|$ for the various estimators. The $\widehat{q}$ above, with $\varepsilon=0.05$, is for example much better than the $\widetilde{q}$ ones, for $q$ in $(0.20,1)$.

(ii) Suppose $a$ comes from $\pi_0(a)$ with probability $p_0$ and from $\pi_1(a)$ with probability $p_1$. Then $\widehat{a}(z)$ can be shown to be of the mixture form $w_0(z)\widehat{a}_0(z)+w_1(z)\widehat{a}_1(z)$, where $\widehat{a}_j(z)$ is the Bayes estimator under theory $p_j(a)$, and $w_j(z)=p_jh_j(z)/\{p_0h_0(z)+p_1h_1(z)\}$, and $h_j(z)=\int\phi(z-a)\pi_j(a)\,da$. An interesting special case is the prior distribution $a\sim(1-\varepsilon)I_0+\varepsilon N\{0,\sigma^2\}$, where $I_0$ denotes the degenerate distribution at zero. This is a 'Bayesian epsilon' approach, where the statistician is rather convinced of the narrow model's correctness but allows the data to express a different opinion with probability $\varepsilon$. In this case

$$\widehat{a}(z)=\frac{\varepsilon}{\varepsilon+(1-\varepsilon)B(z)}\frac{\sigma^2}{\sigma^2+1}z,\quad B(z)=\frac{h_0(z)}{h_1(z)}=\sqrt{\sigma^2+1}\exp\Big\{-\frac{1}{2}\frac{\sigma^2}{\sigma^2+1}z^2\Big\}. \tag{5.12}$$

Again $\sigma^2$ has to be specified or estimated. One possibility is $\widehat{\sigma}^2=Z^2/\varepsilon$, since $Ea^2=\varepsilon\sigma^2$ and $Z$ estimates $a$; other versions can be constructed as in (i) above.

(iii) If it is assumed that $|a|\le m$ a priori then the Bayes solution (5.10) with a uniform prior on $[-m,m]$ should give an estimator with good risk properties on this interval.

*5G. Minimax type estimators.* The remarks about the $a$ parameter in 5D suggest that its range could usefully be taken to be bounded, a priori, in some situations. If $a$ is postulated to be in $[-m,m]$, for example, then estimators $a^*$ exist that are uniformly better than $z$, which means, by our basic correspondence theorem, that estimators $\mu^*$ exist that are uniformly better than $\widehat{\mu}_{\rm wide}$. If in particular $a_m^*$ is a minimax estimator, with maximum risk $r_m<1$ for $R(a)$ in $[-m,m]$, then $\mu^*$ of (5.4), defined via (5.8), has a minimax property: It minimises the limit distribution version of

$$\max_{|\gamma-\gamma_0|\le m\kappa/\sqrt{n}} nE_{\theta_0,\gamma}\{\widehat{\mu}-\mu(\theta_0,\gamma)\}^2$$

over all estimators $\widehat{\mu}$, and achieves $\max_{|\delta|\le m\kappa} E\Lambda^2 = b^2\kappa^2 r_m + \tau_0^2 < \tau^2$.

How do such minimax estimators look like? It is known that $a_m^*(z)$ is the proper Bayes solution w.r.t. a prior distribution concentrated in a finite number of points, see e.g. Lehmann (1983, Chapter 4.3). This least favourable prior has been found for small values of $m$, at least for $m \le 1.5$, and Bickel (1981) gives approximate results for $m$ large. We mention that $a^* = m\tanh(mz)$, the Bayes solution under a symmetric two-point prior in $\pm m$, is minimax, provided $m \le 1.05$. This is relevant here since $[-1,1]$ is the range for $a$ where narrow estimation is better than wide estimation. Bickel shows that the distribution with density $\pi_m(a) = \cos^2(\frac{1}{2}\pi a/m)/m$ for $|a| \le m$ is approximately least favourable, for large $m$. This suggests trying out

$$\widehat{a}_{\rm bic}(z) = \frac{\int_{-m}^{m} a\phi(z-a)\cos^2(\frac{1}{2}\pi a/m)\,da}{\int_{-m}^{m} \phi(z-a)\cos^2(\frac{1}{2}\pi a/m)\,da}, \tag{5.13}$$

It is *not* approximately minimax on $[-m,m]$, but it is uniformly better than $\widehat{a}_{\rm wide} = z$ in a certain interval around 0. A simpler possibility is to use the ML solution when $|a| \le m$ a priori, that is,

$$\widehat{a}_{\rm res}(z) = -m \text{ when } z \le -m, \quad z \text{ on } [-m,m], \quad m \text{ when } z \ge m. \tag{5.14}$$

This is not quite as good as using the proper minimax solution on $[-m,m]$, however.

Finally we should include estimators of the Efron–Morris variety, see Efron and Morris (1971) and Lehmann (1986, Chapter 4.2). These aim at minimising Bayes risk, under normal priors, subject to having maximum risk less than some prescribed level. A particular case of these is pertinent here, namely the 'limited translation estimator'

$$\widehat{a}_{\rm em}(z) = z + m \text{ when } z \le -m, \quad 0 \text{ on } [-m,m], \quad z - m \text{ when } z \ge m. \tag{5.15}$$

These come close to minimising maximum risk subject to doing well at $a = 0$, see Bickel (1983, 1984) and Berger (1982). They are not smooth enough to be admissible. An alternative estimator which can be proposed is

$$\widehat{a}_{\rm atan}(z) = z - m(2/\pi)\arctan z. \tag{5.16}$$

It is motivated from Bickel's study of $\widehat{a}_{\rm em}$ and its connection to bounded influence functions in robust estimation of location, and is scaled so that it has the same maximum risk $1+m^2$ as (5.15) (see below).

*5H. Some concluding comments.*

(i) Observe the generality under which the comparisons of Sections 5 and 6 are made. They are valid and relevant for all of Examples A–G (with appropriate modifications for case B, see Hjort (1991a)) and for all parameter estimands, via (5.4)–(5.7).

(ii) When applied to a particular estimand in a particular model these comparisons should perhaps also include the nonparametric contender. In Example A, for example, one could compare the wide and the narrow parametric methods to the sample median.

(iii) We have been motivated by approximate mean squared error $E_{P_n}(\mu^* - \mu_{\rm true})^2$ when comparing estimator performance. We haven't quite worked with the limit of $n$ times the mean squared error, but rather with $E\Lambda^2$ in (5.6), using the limit distribution. This is both easier and more meaningful. This is a minor technical point, however; usually the two agree. See Lehmann's (1983) Lemma 5.1.2, for example.

(iv) One might wish to study $L_1$ error $\sqrt{n}E_{P_n}|\mu^* - \mu_{\rm true}|$ and its limit distributiuon version $E|\Lambda|$ instead. There is a parallel result to (5.6) and (5.7) for this problem. Let $L(x)$ be the function $E|x + N\{0,1\}| = x + 2\phi(x) - 2x\{1 - \Phi(x)\}$. Then $\mu^*$ of (5.4) has

$$E|\Lambda| = \tau_0 \int L\big((b\kappa/\tau_0)\,(\delta/\kappa - c(z)z)\big)\,\phi(z-a)\,dz = \tau_0 E_a L\big(\rho(c(Z)Z - a)\big), \qquad (5.17)$$

letting $a = \delta/\kappa$ again and $\rho = |b|\kappa/\tau_0$. There is once more a correspondence between compromise estimators $\mu^*$ of $\mu$ and estimators $\widehat{a}(z) = c(z)z$ of $a$, but the $L_1$ loss function $|\mu^* - \mu|$ for $\mu$ is transformed to loss function $L(\rho(\widehat{a} - a))$ for $a$. And there is still a 'tolerance radius' around the narrow model inside of which misspesification is favourable, but one does not get the clear-cut $|\delta| \le \kappa$ answer. The narrow and the wide procedures have respectively $E|b\delta + \tau_0 N|$ and $(b^2\kappa^2 + \tau_0^2)^{1/2}E|N|$ as limiting risks, where $N \sim N\{0,1\}$. The tolerance radius becomes in fact $|\delta| \le a_0\kappa = a_0(\rho)\kappa$, or $|a| \le a_0(\rho)$, where $|a| \le a_0$ corresponds to $E|\rho a + N| \le (1+\rho^2)^{1/2}\,E|N|$. Computations show that $a_0(\rho)$ starts at 1.00 for $\rho = 0$ and slouches towards $\sqrt{2/\pi} = 0.7979$ as $\rho$ grows. The $L_1$ criterion for estimation of $\mu$ accordingly tolerates slightly less misspecification than the $L_2$ criterion.

(v) One might generalise most of this section's results to deviances from the basic model in more than one direction. One can envisage useful generalisations of (5.4), for example, where the final estimator gives weights to the narrow model and to several wider alternative models, with weights determined by the data. Statistics tradition does perhaps dictate this point of view, with a classic null model and several possible departures from it, but the problem can also be turned inside out, starting with a wide a priori model for the data and then smoothing in several directions downwards to narrower models of interest. The empirical Bayes ideas above should be of value if these kind of questions are to be pursued.

## 6. Grand comparison

Each estimator of $\mu_{\rm true}$ has a cousin that estimates $a$ in the $Z \sim N\{a,1\}$ situation, and vice versa, by (5.8). Furthermore, the performance of one of them determines and is determined by the performance of the other one, by the key correspondence (5.6)–(5.7). It is refreshing to judge a $\mu$-estimator by examining its $a$-estimator cousin. Here is a partial list of interesting estimators for $\mu_{\rm true}$, following the various suggestions of Section 5, along with brief descriptions of their performance.

(i) The narrow estimator $\widehat{\mu}_{\rm narr}$ has $c(z) \equiv 0$ and $\widehat{a}(z) \equiv 0$. This particular estimator of $a$ is fully confident that $a$ is close to zero, and has risk $R_{\rm narr}(a) = a^2$.

(ii) The wide estimator $\widehat{\mu}_{\rm wide}$ on the other hand has $c(z) \equiv 1$ and $\widehat{a}(z) = z$. This conservative estimator has constant risk $R_{\rm wide}(a) = 1$, and is the unique admissible minimax estimator for $a$ when the parameter range is unrestricted. Note anew that the narrow is better than the wide when $|a| \le 1$.

(iii) The if-else estimator (5.1), with $m^2$ instead of $1.645^2$ as cut-off point, has $c(z) = I\{|z| \geq m\}$, and corresponds to the $a$-estimator $\widehat{a}(z) = zI\{|z| \geq m\}$. A determined mind finds

$$\begin{aligned} R_{\rm pre}(a) &= \int_{|z|\geq m} (z-a)^2\phi(z-a)\,dz + \int_{|z|\leq m} (0-a)^2\phi(z-a)\,dz \\ &= 1 + (a^2-1)\{\Phi(m+a) + \Phi(m-a) - 1\} \\ &\qquad + (m+a)\phi(m+a) + (m-a)\phi(m-a). \end{aligned}$$

The if-else with cut-off $m = 1$, which corresponds to the Akaike strategy, see 4D, seems overall to be preferable to the one with $m = 1.645$, corresponding to the 10% level test. The latter is better in the vicinity of the narrow model, for $|a| \leq 0.83$, but then becomes markedly worse than the former. The pre-test estimators are not smooth enough to be Bayes or extended Bayes, see (5.10). In particular such methods are not admissible, i.e. they can be improved upon uniformly in $a$! Note that $m = 0$ and $m = \infty$ give back the wide and the narrow methods, respectively. These extreme cases *are* admissible, however.

(iv) The linear combination estimator $\widehat{\mu}_{\rm lin}$ discussed in 5B has $c(z) = c$ and $\widehat{a}(z) = cz$. Its risk is $R_{\rm lin}(a) = c^2 + (1-c)^2a^2$, which is unbounded when $|a|$ grows. These are proper Bayes solutions for $0 \leq c < 1$ and admissible for $0 \leq c \leq 1$.

(v) The natural $\widehat{\mu}_{\rm eb}$ of (5.3) has $c(z) = z^2/(1+z^2)$, and the correspondence to the empirical Bayes estimator $\widehat{a}_{\rm eb}(z) = \{z^2/(1+z^2)\}z$ has been noted in 5F. One must compute

$$R_{\rm eb}(a) = E_a\Big\{\frac{Z^3}{1+Z^2} - a\Big\}^2 = \int\Big\{\frac{z^3}{1+z^2} - a\Big\}^2\phi(z-a)\,dz$$

by numerical integration. I can prove that $\widehat{a}_{\rm eb}$ is admissible. This translates into an admissibility property for $\widehat{\mu}_{\rm eb}$. I have also studied the similarly inspired $\widehat{a}(z) = \widehat{q}(z)z$, with $\widehat{q}(z)$ as in (5.11) instead of $z^2/(z^2+1)$. These perform similarly. The risk for $\widehat{q}(z)z$ starts at 0.37 for $a = 0$, smaller than 0.46 for $\widehat{a}_{\rm eb}(z)$, and stays better for $|a| \leq 1.83$, after which $\widehat{a}_{\rm eb}(z)$ takes over. The maximum risk 1.476 for $\widehat{q}(z)z$ is higher than 1.252 for $\widehat{a}_{\rm eb}(z)$. The risk for $\widehat{a}_{\rm eb}(z)$ is less than the crucial value 1 for $|a| \leq 1.40$. Overall one would argue that $\widehat{a}_{\rm eb}(z)$ is better than $\widehat{q}(z)z$; see also the figure below.

(vi) The restricted ML estimator (5.14) can be shown to have risk function

$$\begin{aligned} R_{\rm res}(a) &= \Phi(m-a) + \Phi(m+a) - 1 - (m-a)\phi(m-a) - (m+a)\phi(m+a) \\ &\qquad + (m-a)^2\{1-\Phi(m-a)\} + (m+a)^2\{1-\Phi(m+a)\}. \end{aligned}$$

This estimator is not smooth enough to be Bayes or extended Bayes, and is like the if-else estimator not admissible. Its risk is satisfactory on $[-m,m]$ but ends up growing as $a^2$ outside it.

(vii) The Efron–Morris estimator (5.15) has risk function

$$R_{\rm em}(a) = 1+m^2+(a^2-m^2-1)\{\Phi(m+a)+\Phi(m-a)-1\}-(m-a)\phi(m+a)-(m+a)\phi(m-a).$$

These increase with $|a|$, and rather rapidly, from a small value at zero towards maximum risk $1+m^2$. The arctan-estimator (5.16) has also risk that increases in $|a|$ from $R_{\rm atan}(0)$ to $1+m^2$. It has higher risk than (5.15) has at $a = 0$, but the risk climbs much more slowly towards $1+m^2$; see also the figure below. In particular an arctan-estimator can be better than an Efron–Morris estimator on $[-5,5]$, say.

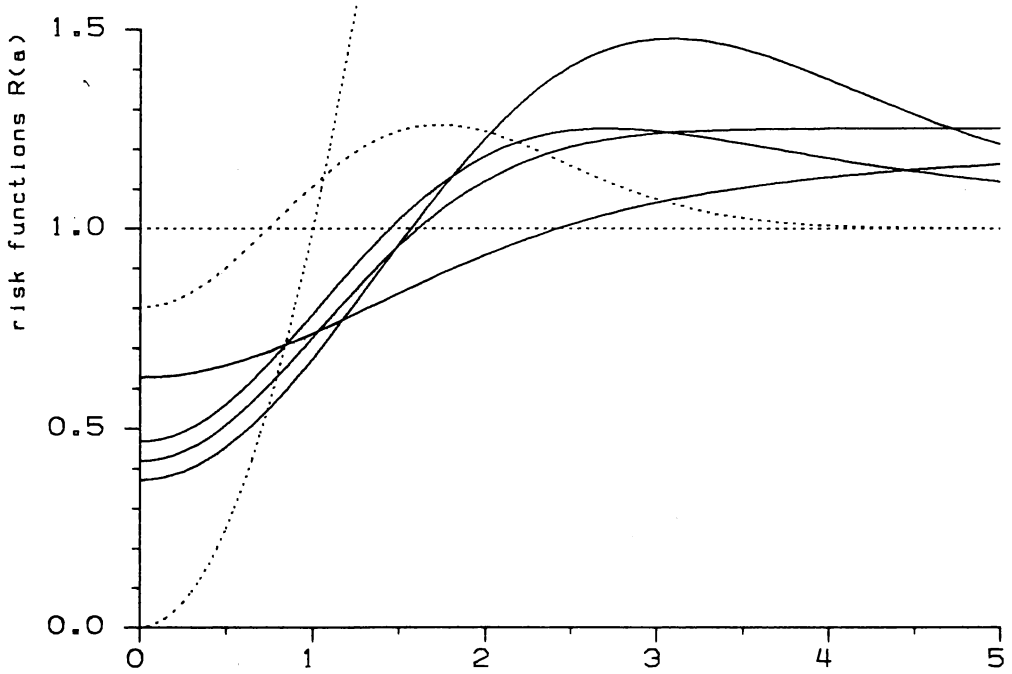


FIGURE. *Risk functions $R(a)$ are shown for seven procedures, corresponding to seven choices of $c(Z_n)$ in (5.4). Risks for the wide and the narrow methods start at 1.00 and 0.00, and are shown with dotted lines, as is the risk for Akaike, which starts at 0.80. The empirical Bayes methods $\widehat{a}_{\rm eb}(Z)$ and $\widehat{q}(Z)Z$ start at 0.47 and 0.37. Finally the Efron–Morris and arctan estimators, both scaled to have the same maximum risk 1.252 as has $\widehat{a}_{\rm eb}(Z)$, start at 0.42 and 0.63.*

(viii) The Bickel-inspired estimator (5.13) has acceptable risk below 1 in an interval around 0, but the risk explodes when $|a|$ grows. The same goes for the Bayes solution with a uniform prior on $[-m, m]$. Its risk is 1 at 0 and at $m$ and below 1 in between, but quickly explodes when $|a|$ grows outside the interval. Note that for $m$ large this solution becomes simply the wide solution.

(ix) The 'epsilon Bayes' methods described in (5.12) and the remarks following it have small risk for $|a|$ less than about 1, but then become markedly worse than both $\widehat{a}_{\rm eb}(Z)$, $\widehat{q}(Z)Z$, and pre-test estimators. The empirical epsilon Bayes method with $\widehat{\sigma}^2 = Z^2/\varepsilon$ is not as good as the simple specified one with $\sigma$ put equal to 3, for example.

Let us compare $\widehat{\mu}_{\rm narr}$, $\widehat{\mu}_{\rm wide}$, the if-else $\widehat{\mu}_{\rm pre}$ with $m = 1$, and the mixture estimator $\widehat{\mu}_{\rm eb}$ of (5.3). The narrow estimator wins if $|a| \leq 0.84$; the mixture estimator wins when $|a|$ is between 0.84 and 1.45, and finally the safest and wide estimator wins if $|a|$ exceeds 1.45. While $\widehat{\mu}_{\rm narr}$ can misbehave significantly when $|\delta| \geq 2.50\kappa$, say, $\widehat{\mu}_{\rm eb}$ always behaves wisely, also in the $|\delta| > \kappa$ case, and does not ever lose much to $\widehat{\mu}_{\rm wide}$. Its worst risk value is 1.252, at $|a| = 2.70$, and when the narrow model is very wrong ($|\delta|$ is large) $\widehat{\mu}_{\rm eb}$ becomes equivalent to $\widehat{\mu}_{\rm wide}$. In no region does $\widehat{\mu}_{\rm pre}$ win, but its risk function lies between the wide method's 1 and the mixture method's risk function, for $|a| > 2.17$; see the figure.

Based on these observations five of the more interesting estimators are singled out for display, in addition to the extreme basis choices 'narrow' and 'wide'. The five are the empirical Bayes versions $\widehat{a}_{\rm eb}(Z)$ and $\widehat{q}(Z)Z$; the Akaike strategy $\widehat{a}_{\rm pre}(Z)$, see 4D and 5A; the Efron–Morris (5.15) with $m = 0.502$ chosen so as to get the same maximum risk

1.252 as $\widehat{a}_{\rm eb}(Z)$; and the smoother arctan-estimator (5.16) with the same $m$ (and the same objective). The Akaike strategy is about as good as the pre-test method can be, but is not as good as the others. It is included since versions of it are in frequent use.

All in all the best choices seem to be the simple empirical Bayes, the Efron–Morris, and the arctan. There are several other methods among those discussed that would make a good show on $[-5,5]$, say, but with risks that explode for growing $|a|$. The $\widehat{\mu}_{\rm eb}$ of (5.3) in particular is a practical and satisfactory solution. There is no artificial cut-off, its weight in favour of the wide model is smoothly increasing from 0 to 1 with the test indicator $Z_n$, it behaves considerably better than the wide estimator in a reasonable neighbourhood of the narrow model, and its maximum risk is only $(1.119\,b\kappa)^2+\tau_0^2$, compared to $(b\kappa)^2+\tau_0^2$ for the conservative wide method. The Efron–Morris and the arctan estimators have similar performances but require selection of a parameter, related to the trade-off between behaving well around zero and having a small maximum risk.

To illustrate more concretely what these suggestions amount to, consider logistic regression as in Example F. If deviation from $\alpha+\beta x$ in direction of quadraticity is suspected, use

$$p^*(x)=\frac{1}{1+Z_n^2}\frac{\exp(\widehat{\alpha}_{\rm narr}+\widehat{\beta}_{\rm narr}x)}{1+\exp(\widehat{\alpha}_{\rm narr}+\widehat{\beta}_{\rm narr}x)}+\frac{Z_n^2}{1+Z_n^2}\frac{\exp(\widehat{\alpha}+\widehat{\beta}x+\widehat{\gamma}x^2)}{1+\exp(\widehat{\alpha}+\widehat{\beta}x+\widehat{\gamma}x^2)},$$

where $Z_n=\sqrt{n}\widehat{\gamma}/\widehat{\kappa}$. Or replace the weights with $1-\widehat{a}(Z_n)/Z_n$ and $\widehat{a}(Z_n)/Z_n$, with $\widehat{a}(Z_n)$ equal to the limited translation estimator (5.15) or the arctan-estimator (5.16).

The facts above are meant to summarise the main features of the various estimator performances, based on a thorough investigation and several days of conscentious staring at hundreds of risk functions. Computer programs and risk tables for each of the estimator classes discussed above are collected in Hjort (1991b), which is available upon courteous request.

## 7. Examples

We now provide answers to the questions asked in Examples A–G of the introduction!

EXAMPLE A. In the general two-parameter Weibull model, parameterised as in (1.1), the score function becomes

$$\begin{pmatrix}\frac{\gamma}{\theta}\{1-(\theta y)^\gamma\}\\ \frac{1}{\gamma}\{1+\log(\theta y)^\gamma-(\theta y)^\gamma\log(\theta y)^\gamma\}\end{pmatrix},$$

and clever computations involving the gamma function and its derivatives reveal the information matrix and its inverse to be

$$J_{\rm gen}=\begin{pmatrix}\gamma^2/\theta^2 & (1-k)/\theta\\ (1-k)/\theta & c^2/\gamma^2\end{pmatrix},\quad J_{\rm gen}^{-1}=\frac{1}{\pi^2/6}\begin{pmatrix}c^2\theta^2/\gamma^2 & -(1-k)\theta\\ -(1-k)\theta & \gamma^2\end{pmatrix},$$

in which $k=0.577...$ is the Euler–Mascheroni constant and $c^2=\pi^2/6+(1-k)^2$. The null model corresponds to $\gamma=\gamma_0=1$. The $\kappa^2$ parameter is $6/\pi^2$, and we have reached the following conclusion: For $|\gamma-1|\le\sqrt{6/\pi^2}/\sqrt{n}=0.779/\sqrt{n}$, estimation with $\mu(1/\bar{Y},1)$ based on simple and narrow-minded exponentiality performs better than high-brow $\mu(\widehat{\theta},\widehat{\gamma})$; and this is true regardless of the parameter $\mu$ to be estimated.

In the language of 4F Weibull deviance from exponentiality has tolerance limit $d = J_{22}J^{22} = 1+(1-k)^2/(\pi^2/6) = 1.109$, and $\rho^2$ of (4.3) is $(1-k)^2/\{(1-k)^2+\pi^2/6\} = 0.098$. It is instructive to compare these with corresponding values for gamma distribution deviance from exponentiality. If $f(y) = \{\theta^\gamma/\Gamma(\gamma)\}\, y^{\gamma-1}e^{-\theta y}$ is the gamma density, for which $\gamma_0 = 1$ gives back exponentiality, then $\kappa^2 = 1/(\pi^2/6-1)$; estimation using $\mu(1/\bar{Y},1)$ is more precise than $\mu(\widehat{\theta},\widehat{\gamma})$ provided $|\gamma-1| \leq 1.245/\sqrt{n}$; $d$ is 2.551; and $\rho^2 = 6/\pi^2 = 0.608$. This suggests that moderate gamma-ness is less critical than moderate Weibull-ness for standard methods based on exponentiality.

EXAMPLE B. The wide model has parameters $\xi$, $\sigma$, $m$. Let us reparameterise to $\gamma = 1/m$, so that the density becomes

$$f(y,\xi,\sigma,\gamma) = \frac{c(\gamma)}{\sigma}\Big\{1+\gamma\Big(\frac{y-\xi}{\sigma}\Big)^2\Big\}^{-\{1/2+1/(2\gamma)\}}, \quad c(\gamma) = \frac{\sqrt{\gamma}}{\sqrt{\pi}}\frac{\Gamma(1/2+1/(2\gamma))}{\Gamma(1/(2\gamma))}.$$

Estimation of the model parameters must now be studied when $\gamma$ is small and nonnegative. This actually calls for special treatment since the null point $\gamma = 0$ is not an inner point, and $\widehat{\gamma} = 0$, or $\widehat{m} = \infty$, happens with positive probability. Such a treatment is given in Hjort (1991a), and shows that if $\gamma \leq 0.686/\sqrt{n}$, i.e. if the degrees of freedom $m \geq 1.458\sqrt{n}$, then $t$-ness doesn't matter, and *any* parameter $\mu = \mu(f) = \mu(\xi,\sigma,m)$ is better estimated in the ordinary, simple, normality based way. A similar result is also proven there for regression models.

EXAMPLE C. We generalise slightly and write the wide model as $Y_i \sim N\{\beta' x_i + \gamma c(x_i), \sigma^2\}$, where $\beta$ and $x_i$ are $p$-dimensional vectors, and $c(x)$ is some given scalar function. By computing log-derivatives and evaluating covariances one reaches

$$J_{n,\text{wide}} = \frac{1}{\sigma^2}\begin{pmatrix} 2 & 0 & 0 \\ 0 & n^{-1}\sum_{i=1}^n x_i x_i' & n^{-1}\sum_{i=1}^n x_i c(x_i) \\ 0 & n^{-1}\sum_{i=1}^n x_i' c(x_i) & n^{-1}\sum_{i=1}^n c(x_i)^2 \end{pmatrix},$$

from the definition in (3.4). It follows that

$$\kappa^2 = \sigma^2 \times \text{lower right element of } \begin{pmatrix} n^{-1}\sum_{i=1}^n x_i x_i' & n^{-1}\sum_{i=1}^n x_i c(x_i) \\ n^{-1}\sum_{i=1}^n x_i' c(x_i) & n^{-1}\sum_{i=1}^n c(x_i)^2 \end{pmatrix}^{-1}.$$

Assume, for a concrete example, that $x_i$ is one-dimensional and uniformly distributed over $[0,b]$, say $x_i = b\frac{i}{n+1}$, and that the wide model has $\alpha+\beta(x_i-\bar{x})+\gamma(x_i-\bar{x})^2$. Then $\kappa \doteq \sqrt{80}\,\sigma/b^2$. Consequently, dropping the quadratic term does not matter, and is actually advantageous, for every estimator, provided $|\gamma| \leq 8.94\,\sigma/b\sqrt{n}$. In many situations with moderate $n$ this will indicate that it is best to keep the narrow model and avoid quadratic analysis.

EXAMPLE D. Again we are mildly general and write $Y_i \sim N\{\beta' x_i, \sigma^2(1+\gamma c(x_i))\}$ for the $p+2$ parameter variance heterogeneous model. It is not easy to put up simple expressions for the general information matrix, in the presence of $\gamma$, but once more we are

permitted to compute $J_{n,\text{wide}}$ and $J_{\text{wide}}$ of (3.4) under the null model, that is, when $\gamma = 0$. Some calculations give

$$J_{n,\text{wide}} = \begin{pmatrix} 2/\sigma^2 & 0 & \sigma^{-1}n^{-1}\sum_{i=1}^n c(x_i) \\ 0 & \sigma^{-2}n^{-1}\sum_{i=1}^n x_i x_i' & 0 \\ \sigma^{-1}n^{-1}\sum_{i=1}^n c(x_i) & 0 & (2n)^{-1}\sum_{i=1}^n c(x_i)^2 \end{pmatrix}.$$

Matters simplify and $1/\kappa^2$ is found to be $(2n)^{-1}\sum_{i=1}^n \{c(x_i) - \bar{c}\}^2$. If once again $x_i$'s are distributed evenly on $[0, b]$, and $c(x_i) = x_i$, then $\kappa \doteq \sqrt{24}/b$, and the criterion becomes $|\gamma| \leq 4.90/b\sqrt{n}$. In particular this shows that the sophisticated variance heterogeneous approach, which uses the weighted least squares estimator

$$\widehat{\beta}_{\text{soph}} = \sum_{i=1}^n \frac{x_i y_i}{1 + \widehat{\gamma} c(x_i)} \Big/ \sum_{i=1}^n \frac{x_i^2}{1 + \widehat{\gamma} c(x_i)},$$

is inferior to the simpler solution, unless $|\gamma|$ is quite large. It is of course the sampling variability present in the weights, via the ML estimator $\widehat{\gamma}$, that makes $\widehat{\beta}_{\text{soph}}$ inferior to ordinary $\widehat{\beta}_{\text{narr}}$.

EXAMPLE E. Assume that a $h_\lambda$-transformation of $(Y_i - \beta' x_i)/\sigma$ is $N\{0,1\}$. When $h_\lambda(Z)$ is $N\{0,1\}$, then $Z$ has cumulative $\Phi(z)^\lambda$ and density $\lambda\Phi(z)^{\lambda-1}\phi(z)$. Hence $Y_i$ has density

$$f(y_i, \sigma, \beta, \lambda | x_i) = \lambda \Phi\big((y_i - \beta' x_i)/\sigma\big)^{\lambda-1} \phi\big((y_i - \beta' x_i)/\sigma\big)/\sigma.$$

Is it now possible to evaluate partial derivatives w.r.t. $\sigma$, $\beta$, $\lambda$. Their null model versions, corresponding to $\lambda = 1$, become $(z_i^2 - 1)/\sigma$, $z_i x_i/\sigma$, $1 + \log \Phi(z_i)$, where $z_i = (y_i - \beta' x_i)/\sigma$. Formula (3.4) gives the $(p+2) \times (p+2)$ matrix

$$J_{n,\text{wide}} = \begin{pmatrix} 2/\sigma^2 & 0 & b/\sigma \\ 0 & \sigma^{-2}n^{-1}\sum_{i=1}^n x_i x_i' & a\sigma^{-1}n^{-1}\sum_{i=1}^n x_i \\ b/\sigma & a\sigma^{-1}n^{-1}\sum_{i=1}^n x_i & 1 \end{pmatrix},$$

in which $a = EN \log \Phi(N) = 0.9032$ and $b = E\{1 + N^2 \log \Phi(N)\} = -0.5956$ (computed by numerical integration). It follows that

$$1/\kappa^2 = 1 - \tfrac{1}{2}b^2 - a^2 \bar{x}' \Big\{ \frac{1}{n} \sum_{i=1}^n x_i x_i' \Big\}^{-1} \bar{x}.$$

$\kappa$ can be rather large, which in turn means that standard regression copes well even if $\lambda$ differs quite a bit from 1. If only $|\lambda - 1| \leq \kappa/\sqrt{n}$, then standard regression methods work better than cumbersome ones employing a separate estimate for $\lambda$.

In the special case of a constant mean the tolerance limit against misspecification is very relaxed, with $\kappa = 12.090$. In this case $V = 1 + \log \Phi(z)$ is extremely well explained by $U = (z/\sigma, (z^2 - 1)/\sigma)$, with a maximal correlation of 0.993; see the discussion under 4F. The classic $N\{\xi, \sigma^2\}$ can stand a good deal of misspecification w.r.t. $\lambda$. — In another special case, that of $\alpha + \beta(x_i - \bar{x}) + \sigma Z_i$, $1/\kappa^2$ becomes $1 - \frac{1}{2}b^2$ and $\kappa$ becomes much smaller, namely 1.103. In the language of 4F the $n$ values of $V_i = 1 + \log \Phi(z_i)$ are now

much less well explained by the respective values of $U_i = \big(z_i, (x_i - \bar{x})z_i, z_i^2 - 1\big)/\sigma$, and the standard regression model can only tolerate up to $1.103/\sqrt{n}$ deviance from $\lambda = 1$.

EXAMPLE F. Write $p_i = p(x_i, \beta, \gamma)$, in which $\gamma = \gamma_0$ gives back ordinary logistic regression, and write $p_i^0$ for $p(x_i, \beta_0, \gamma_0)$ at some target point $\beta_0$. It is not difficult to reach

$$J_{n,\text{wide}} = \frac{1}{n}\sum_{i=1}^n \frac{1}{p_i^0(1-p_i^0)} \begin{pmatrix} \partial p_i/\partial\beta \\ \partial p_i/\partial\gamma \end{pmatrix} \begin{pmatrix} \partial p_i/\partial\beta \\ \partial p_i/\partial\gamma \end{pmatrix}',$$

where the partial derivates are computed at the null point as usual. Finding the tolerance limit $\kappa^2$ is achieved by computing this $(p+1) \times (p+1)$ matrix numerically, at the target point, which will typically be the estimated $\widehat{\beta}_{\text{narr}}$ computed from ordinary analysis, and then inverting it; $\kappa^2$ is found at the lower right corner.

In the two types of model departure discussed in Section 1, this goes as follows. If the wide model says $\alpha + \beta(x_i - \bar{x}) + \gamma(x_i - \bar{x})^2$, then

$$J_{n,\text{wide}} = \frac{1}{n}\sum_{i=1}^n p_i^0(1-p_i^0) \begin{pmatrix} 1 & t_i & t_i^2 \\ t_i & t_i^2 & t_i^3 \\ t_i^2 & t_i^3 & t_i^4 \end{pmatrix},$$

where $t_i = x_i - \bar{x}$. In the case of (1.5), on the other hand, involving a shape parameter $\eta$,

$$J_{n,\text{wide}} = \frac{1}{n}\sum_{i=1}^n \frac{p_i^0}{1-p_i^0} \begin{pmatrix} (1-p_i^0)^2 & (1-p_i^0)^2 x_i & (1-p_i^0)\log p_i^0 \\ (1-p_i^0)^2 x_i & (1-p_i^0)^2 x_i^2 & (1-p_i^0)\log p_i^0\, x_i \\ (1-p_i^0)\log p_i^0 & (1-p_i^0)\log p_i^0\, x_i & (\log p_i^0)^2 \end{pmatrix}.$$

EXAMPLE G. Write $\sigma_1^2 = \sigma^2$ and $\sigma_2^2 = \sigma^2(1+\gamma)$. Finding the $J_{\text{wide}}$ matrix in the $(\xi_1, \xi_2, \sigma, \gamma)$ model is not difficult, and leads to $\kappa^2 = 2/\{r(1-r)\}$, where $r = m/(m+n)$. This means a tolerance level of $d = \kappa^2 J_{22} = 1/r = (m+n)/m$. The simple equal variance model can tolerate $\gamma^2 \le 2(m+n)/mn$, which becomes $|\gamma| \le 2/\sqrt{n}$ in the $m = n$ case. This is a fairly low tolerance limit, and different variances qualifies as a dangerous departure from the narrow model.

OTHER EXAMPLES. There is a large variety of other examples of common departures from standard models and that could be studied using our general methods and results. In each case one could compute the tolerance radius, one could speculate about robustness against the deviation in question in light of $d$ and $\rho$ of 4F, and one could implement the method of (5.3), for example. A partial list of such models and deviations is: (i) Typical i.i.d. models against various forms of dependence. (ii) Multinomial and log-linear models against higher order interactions. (iii) Analysis of variance models against interaction terms. (iv) Analysis of variance models against different variances in different groups. (v) Regression models against presence of cross-terms. (vi) Time series models against higher order autoregression or moving average. (vii) Typical i.i.d. models for discrete variables against Markov dependence. (viii) Markov chain models against second order Markovness. (ix) Models with normal errors against contamination of gross errors. (x) Traditional homogeneous models in survival analysis against heterogeneous frailness of individuals.

## References


Berger, J.O. (1982). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. In *Statistical Decision Theory and Related Topics III*, eds. Berger and Gupta, 109–141. Academic Press, New York.

Bickel, P.J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.* **9**, 1301–1309.

Bickel, P.J. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. *Recent Advances in Statistics, Festschrift for Herman Chernoff*, eds. Rizvi and Siegmund, 511–528. Academic Press, New York.

Bickel, P.J. (1984). Parametric robustness: small biases can be worthwile. *Ann. Statist.* **12**, 864–879.

Efron, B. and Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators. *J. Amer. Statist. Assoc.* **66**, 807–815.

Hjort, N.L. (1986a). *Theory of Statistical Symbol Recognition.* Research monograph, Norwegian Computing Centre, Oslo.

Hjort, N.L. (1986b). Bayes estimators and asymptotic efficiency in parametric counting process models. *Scand. J. Statist.* **13**, 63–85.

Hjort, N.L. (1991a). The exact amount of t-ness that the normal model can tolerate. Technical report, University of Oslo; submitted for publication.

Hjort, N.L. (1991b). Computer programs and risk functions for estimators for a normal mean. Technical report, University of Oslo; available upon courteous request.

Lehmann, E.L. (1983). *Theory of Point Estimation.* Wiley, Singapore.